Elisabete
Dos Santos Ferreira

Métodos Biplot aplicados a dados de Biologia Molecular

Elisabete
Dos Santos Ferreira

Métodos Biplot aplicados a dados de Biologia Molecular

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica da Dra. Adelaide de Fátima Baptista Valente Freitas, Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro.

Ao meu namorado e minha família

o júri / the jury

presidente / president

Prof. Dra. Isabel Maria Simões Pereira
Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro

vogais / examiners committee

Prof. Dra. Adelaide de Fátima Baptista Valente Freitas
Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro (orientador)

Prof. Dra. Vera Mónica Almeida Afreixo
Professor Auxiliar Convidada do Departamento de Matemática da Universidade de Aveiro (co-orientador)

Prof. Dr. Valter Martins Vairinhos
Professor Auxiliar do Instituto Superior de Línguas e Administração de Santarém

agradecimentos / acknowledgements

Em primeiro lugar, quero agradecer as minhas orientadoras, Prof. Doutora Adelaide de Fátima Baptista Valente Freitas e Prof. Doutora Vera Mónica Almeida Afreixo, por todo o apoio disponibilizado e empenho ao longo da dissertação de Mestrado, pois sem a ajuda delas, seria impossível concluir este trabalho.

Também quero agradecer ao Doutor Mira Park pelos esclarecimentos fornecidos relativamente às dúvidas que surgiram no estudo do seu artigo que é uma base fundamental desta dissertação.

Aos meus amigos e colegas que me ajudaram e apoiaram nos momentos mais difíceis. Nomeadamente a Eloísa e ao Luís que se mostraram sempre disponíveis para me auxiliar sempre que precisei.

Por último, os meus sinceros agradecimentos, ao meu namorado e família e a minha família por terem acreditado em mim e terem sempre prestado um apoio incansável.

palavras-chave biplot, Análise de Componentes Principais (ACP), Análise de Correspondências (AC), Escalonamento Multidimensional (MDS), Decomposição em Valores Singulares (SVD).

Resumo

Uma análise Estatística Multivariada (EM) surge sempre que os dados de uma determinada amostra incluem medições simultâneas de mais que uma variável. Nesta dissertação usamos EM no sentido da redução da dimensionalidade dos dados. Para isso usamos as técnicas ACP, AC e MDS.

A ACP permite reduzir o conjunto original de variáveis num menor conjunto de variáveis independentes, explicando a máxima variabilidade das variáveis originais. A AC converte uma matriz de dados não negativos num tipo de representação gráfica em que as linhas e colunas da matriz são representadas por pontos no gráfico. O MDS visa representar as distâncias entre pontos num sistema de dimensão reduzida, de modo a que a distância euclidiana entre eles reproduza aproximadamente as "distâncias originais" no conjunto de dados originais.

Para a representação dos dados num espaço reduzido usamos os biplots. Um biplot é um gráfico de dispersão (de dimensão 2 ou 3), que representa simultaneamente marcadores para as variáveis e marcadores para os objectos.

O objectivo deste trabalho é estudar as técnicas ACP, AC, MDS e os biplots, e aplicá-las a duas bases de dados reais (dados sobre o cancro do cólon e sequências completas de DNA de 123 espécies repartidas em 5 reinos) usando o software "R".

Por fim, são apresentados os resultados obtidos.

keywords biplot, Principal Component Analysis (PCA), Correspondence Analysis (CA), Multidimensional Scaling (MDS), Singular Value Decomposition (SVD)

abstract

Multivariate statistical analysis is used when the data of a given sample includes simultaneous measurements on many variables. With the purpose of reducing data dimensionality, multivariate statistical analysis is applied. Three methods are addressed: PCA, CA and MDS.

PCA reduces the original set of variables into a smaller set of independent variables, explaining the maximum variability of the original variables. CA converts a matrix of nonnegative data into a particular type of graphical display in which the rows and columns of the matrix are depicted as points. The objective of MDS is to display graphically the distance between points in a low-dimensional system so that the euclidean distance between them approximately reproduces the "original distances" in the original data set.

For the representation of data in a low-dimensional space we use biplots. A biplot is a scatter plot (of low-dimension, 2 or 3), which display shows column effects of the variables and row effects of the obervations.

The objective of this paper is to study PCA, CA and MDS methods and biplots. We also apply these methods to two real data sets (colon cancer data set and complete sequences of DNA from 123 species divided into 5 kingdoms) using the software "R".

Finally, we present the results obtained.

Conteúdo

1 Introdução **1**

1.1 Biologia Molecular 1

1.2 Estatística Multivariada na Biologia Molecular 6

1.3 Motivação e Organização da Dissertação 8

2 Métodos Biplot **11**

2.1 Conceitos Fundamentais 11

2.2 Redução da dimensionalidade dos dados 18

2.2.1 Análise de Componentes Principais 18

2.2.2 Análise de Correspondências 22

2.2.3 Escalonamento Multidimensional 27

2.2.4 Vantagens e desvantagens 32

2.3 Biplot com redução da dimensionalidade 33

2.3.1 ACP Biplot 33

2.3.2 AC Biplot 40

2.3.3 MDS Biplot 43

3 Estudo Experimental **47**

3.1 Matrizes de dados 47

3.1.1 Dados 1: Dados de *microarrays* 47

3.1.2 Dados 2: Dados de pares de codões nas sequências de DNA 48

3.2 Resultados e Análise 49

3.2.1 Dados 1: Dados de *microarrays* 49

3.2.2 Dados 2: Dados sobre pares de codões nas sequências de DNA 57

4 Conclusões e Trabalhos Futuros **65**

A Comandos usados no R **67**

A.1 Dados *Iris* 67

A.1.1 ACP para os dados *Iris* 67

A.1.2 AC para os dados *Iris* 68

A.1.3 MDS para os dados *Iris* 69

A.2 Dados 1: Dados de *microarrays* 70

A.2.1 ACP para os Dados 1 70

A.2.2 ACP para o conjunto Dados 1 restrito 71

A.2.3 AC para os Dados 1 72

A.2.4 MDS para os Dados 1 74

A.3 Dados 2: Dados sobre pares de codões nas sequências de DNA 74

A.3.1 ACP para os Dados 2 75

A.3.2 ACP para o conjunto de Dados 2 restrito 76

A.3.3 AC para os Dados 2 77

A.3.4 MDS para os Dados 2 79

B Tabelas **81**

Bibliografia **85**

Lista de Figuras

1.1 Estrutura do DNA e RNA (Vieira [50]). 2
1.2 Código genético que estabelece a correspondência entre tripletos de nucleótidos e aminoácidos (da Silva et al. [18]). 3
1.3 Processo de análise de *microarrays*. 5

2.1 Projecção dos pontos na 1ª CP (adaptada de Villardón [52] e Álvarez González [35]). 18
2.2 Gráficos obtidos através da ACP biplot para os dados *Iris*. 39
2.3 Gráficos obtidos através da AC biplot para os dados *Iris*. 43
2.4 Gráficos obtidos através do MDS biplot para os dados *Iris*. 45

3.1 Gráficos obtidos através da ACP para o conjunto Dados 1. 51
3.2 Biplots obtidos através da ACP para o conjunto Dados 1. 52
3.3 Gráficos obtidos através da AC biplot para o conjunto Dados 1. 55
3.4 Gráficos com as contribuições absolutas e relativas, obtidos através da AC biplot para o conjunto Dados 1. 55
3.5 AC biplot sem recurso ao comando `plot(ca(···))` para as coordenadas principais dos indivíduos, com a identificação, para o conjunto Dados 1. 56
3.6 Gráficos obtidos através do MDS biplot para o conjunto Dados 1. 57
3.7 Gráficos obtidos através da ACP para o conjunto Dados 2. 58
3.8 Biplots obtidos através da ACP para o conjunto Dados 2. 60
3.9 Gráficos obtidos através da AC biplot para o conjunto Dados 2. 61
3.10 Gráficos com as contribuições absolutas e relativas, obtidos através da AC biplot para o conjunto Dados 2. 61
3.11 AC biplot sem recurso ao comando `plot(ca(···))` para as coordenadas principais das 123 espécies em análise, com a identificação, para o conjunto Dados 2. 62
3.12 Gráficos obtidos através do MDS biplot para o conjunto Dados 2. 63

Lista de Tabelas

1.1 Tabela resultante do Anaconda. 6

3.1 Tabela representativa do ficheiro de dados relativos ao conjunto Dados 2. . . . 49
3.2 Lista dos indivíduos com número correspondente utilizado na AC. 54

B.1 Lista das 123 espécies em estudo, com a designação de cada espécie. 81

Nomenclatura

$\mathbf{X}_{n\times p}$	matriz de dados com n indivíduos e p variáveis com característica k.
$\mathbf{V}_{p\times p}$	matriz ortogonal dos vectores singulares à direita.
$\mathbf{U}_{n\times n}$	matriz ortogonal dos vectores singulares à esquerda.
$\mathbf{G}_{n\times k}$	matriz representativa dos marcadores dos indivíduos.
$\mathbf{H}_{p\times k}$	matriz representativa dos marcadores das variáveis.
$\mathbf{\Sigma}_{n\times p}$	matriz diagonal que contém os k valores próprios de $\mathbf{X}$.
$\mathbf{S}_{p\times p}$	matriz das covariâncias de $\mathbf{X}$ com valores próprios $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$.
$\mathbf{l}_i^T$	vector de constantes reais de dimensão p.
$\mathbf{D}_{n\times p}$	matriz que resulta da centralização por colunas da matriz de dados $\mathbf{X}$.
s_j	desvio padrão da variável j.
$\overline{x}_j$	média da variável j.
$\mathbf{Y}_{q\times p}$	matriz representativa das novas componentes principais.
$\mathbf{N}_{a\times b}$	matriz de frequências absolutas com dois factores de classificação A e B.
$\mathbf{T}_{a\times b}$	matriz de correspondências.
$\mathbf{r}$	vector de massas de linha de dimensão a.
$\mathbf{c}$	vector de massas de coluna de dimensão b.
$\mathbf{P_L}$	matriz de dimensão $a \times b$ dos perfis de linha.
$\mathbf{P_C}$	matriz de dimensão $a \times b$ dos perfis de coluna.
$\mathbf{Diag}_{\mathbf{r}}^{-1}$	matriz diagonal de dimensão $(a \times a)$ dos valores recíprocos do vector de massas $\mathbf{r}$.
$\mathbf{Diag}_{\mathbf{c}}^{-1}$	matriz diagonal de dimensão $(b \times b)$ dos valores recíprocos do vector de massas $\mathbf{c}$.
$\mathbf{N}(A)$	nuvem de perfis de linha constituída por a pontos em $\mathbb{R}^b$.
$\mathbf{N}(B)$	nuvem de perfis de coluna constituída por b pontos em $\mathbb{R}^a$.
$\mathbf{\Delta}_{n\times n}$	matriz das distâncias obtida através de $\mathbf{X}$.
$\mathbf{I}_{n\times n}$	matriz identidade.
$\mathbf{P}_{\mathbf{1}_n}$	matriz de projecção ortogonal.
$\mathbf{A}_{n\times n}$	matriz transformada das distâncias.
$\mathbf{Q}_{n\times n}$	matriz dos produtos internos.

Capítulo 1

Introdução

1.1 Biologia Molecular

A Biologia Molecular é o estudo da Biologia ao nível molecular com especificidade na herança da informação genética que está contida nos cromossomas e na função dos genes. Os genes codificam a informação necessária para a síntese de proteínas. O ácido desoxirribonucleico (*deoxyribonucleic acid*, DNA) - é considerado como o suporte universal da informação genética.

Segundo Watson e Crick, o DNA é uma longa molécula em forma de dupla hélice, contendo sequências de nucleótidos (ver Figura 1.1). Cada um dos nucleótidos que entra na molécula de DNA tem na sua constituição três componentes: um grupo fosfato (ácido fosfórico, que confere à molécula características ácidas), uma pentose (a desoxirribose $C_5H_{10}O_4$) e uma base azotada (bases pirimídicas e bases púricas). As bases pirimídicas são bases de anel simples, timina (T) e citosina (C) e as bases púricas são bases de anel duplo, adenina (A) e guanina (G). O número e a ordem dos nucleótidos numa sequência de DNA é muito importante, pois contém a informação genética que define as características de cada indivíduo (da Silva et al. [18]).

As bandas laterais da dupla hélice que contêm uma molécula de DNA são formadas pelas moléculas do grupo fosfato, que alternam com moléculas de desoxirribose, e os "degraus" centrais são pares de bases ligados entre si por pontes de hidrogénio. A especificidade de ligações de hidrogénio entre as bases é chamada de complementaridade de bases:

- a adenina liga-se à timina, e
- a guanina liga-se à citosina.

No caso do ácido ribonucleico (*ribonucleic acid*, RNA), a adenina liga-se ao uracilo, pois a timina só existe no DNA e o uracilo só existe no RNA.

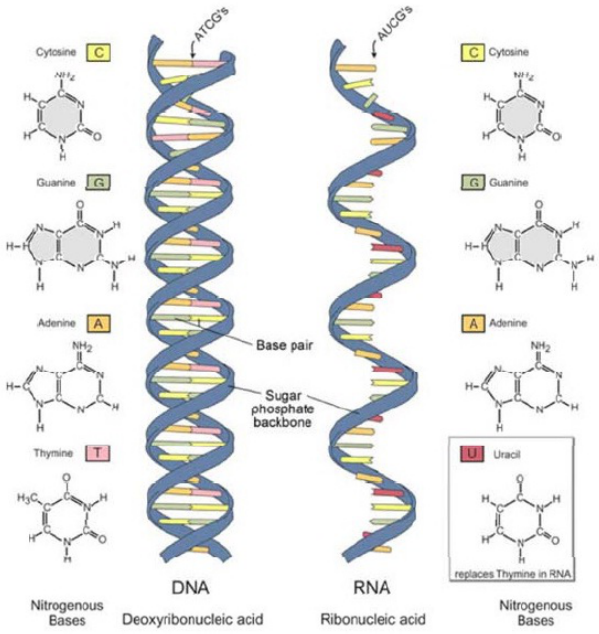


Figura 1.1: Estrutura do DNA e RNA (Vieira [50]).

O DNA tem a capacidade de se autoduplicar, assegurando a conservação do património genético de célula para célula ao longo das gerações. À Biologia Molecular interessa investigar os mecanismos de replicação do DNA assim como a expressão da informação genética. São decifradas as suas características e mecanismos da síntese proteica que permitem transcrever, traduzir, e formar cada proteína.

Estudos efectuados mostraram que o fluxo da informação genética dá-se do DNA para o RNA mensageiro (mRNA), através do processo de transcrição e do processo de tradução. A descoberta do mRNA e a sua importância na síntese proteica conduziu à decifração do código genético (Oliveira et al. [41]).

Os investigadores (na área da Biologia Molecular) estabeleceram um código de correspondência - código genético - usando a linguagem das quatro letras A, C, G e T do DNA, correspondentes aos quatro tipos de nucleótidos. Verificaram que este é o sistema de codificação mais simples utilizado pelas células vivas. Nessa linguagem, três nucleótidos consecutivos do DNA constituem um tripleto o qual representa a mais pequena unidade de mensagem genética necessária à codificação de um aminoácido. Existem $4^3 = 64$ diferentes tipos de tripletos para codificar os vinte e dois possíveis aminoácidos que caracterizam diversas proteínas e cada

tripleto do mRNA, que codifica um determinado aminoácido ou início ou fim de síntese de proteínas, denomina-se codão. A linguagem dos nucleótidos traduz as sequências de codões, e consequentemente dos aminoácidos envolvidos na codificação das proteínas (Figura 1.2).

1ª BASE	2ª BASE: U	2ª BASE: C	2ª BASE: A	2ª BASE: G	3ª BASE
U	UUU, UUC: Fenilalanina (Fen); UUA, UUG: Leucina (Leu)	UCU, UCC, UCA, UCG: Serina (Ser)	UAU, UAC: Tirosina (Tir); UAA: codão de finalização; UAG: codão de finalização	UGU, UGC: Cisteína (Cis); UGA: codão de finalização; UGG: Tiptofano (Trp)	U C A G
C	CUU, CUC, CUA, CUG: Leucina (Leu)	CCU, CCC, CCA, CCG: Prolina (Pro)	CAU, CAC: Histidina (His); CAA, CAG: Glutamina (Glu)	CGU, CGC, CGA, CGG: Arginina (Arg)	U C A G
A	AUU, AUC, AUA: Isoleucina (Ile); AUG: Metionina (Met) codoão de iniciação	ACU, ACC, ACA, ACG: Treonina (Tre)	AAU, AAC: Asparagina (Asn); AAA, AAG: Lisina (Lis)	AGU, AGC: Serina (Ser); AGA, AGG: Arginina (Arg)	U C A G
G	GUU, GUC, GUA, GUG: Valina (Val)	GCU, GCC, GCA, GCG: Alanina (Ala)	GAU, GAC: Ácido aspártico (Asp); GAA, GAG: Ácido glutâmico (Glu)	GGU, GGC, GGA, GGG: Glicina (Gli)	U C A G

Figura 1.2: Código genético que estabelece a correspondência entre tripletos de nucleótidos e aminoácidos (da Silva et al. [18]).

Vários estudos relativos ao código genético permitiram identificar algumas das suas características:

- Universalidade do código genético - Há uma linguagem que é comum a quase todas as células, desde os organismos mais simples aos mais complexos.
- O código não é ambíguo - A um tripleto de nucleótidos corresponde um aminoácido e um só, sempre o mesmo.
- O código genético é redundante - Significa que vários codões são sinónimos, ou seja, podem codificar o mesmo aminoácido.
- O terceiro nucleótido de cada codão não é tão especifico como os dois primeiros - Por exemplo, o aminoácido arginina pode ser codificado pelos codões CGU, CGC, CGA, ou CGG.
- O tripleto AUG tem dupla função - Este tripleto codifica o aminoácido metionina e é também o codão de iniciação da síntese proteica.
- Os tripletos UAA, UAG, UGA são codões de finalização - Estes codões assinalam o fim de síntese proteica e não codificam aminoácidos.

Para além do estudo da estrutura física que define a informação genética, na última década, a Biologia Molecular apresentou grande desenvolvimento com a utilização de uma técnica experimental conhecida por *microarray*. Esta tecnologia tem impulsionado, de maneira importante, a pesquisa da genómica funcional dos diferentes organismos, desde bactérias até ao Homem, incluindo situações normais e patológicas (cancro, doenças auto-imunes, doenças degenerativas entre outras) (sit [10]). É utilizada para identificar genes específicos de determinados tecidos, como também para desenvolver novos fármacos, por exemplo, para diabetes e alguns tipos de cancro (ver referência contida em (Pierce [44])). Na realidade, graças à pesquisa genómica, é possível entender como os genes influenciam o aparecimento de certas doenças. A sua análise torna-se fulcral para criar diagnósticos e medicamentos eficientes para o tratamento em questão.

A técnica dos *microarrays* procura medir os níveis de expressão de milhares de genes simultaneamente, permitindo o estudo da função dos genes em larga escala e assim obter uma perspectiva da expressão relativa de todos os genes de um dado organismo ou tecido.

Um *microarray* de DNA, também designado por chip de DNA, consiste num *array* prédefinido de moléculas de DNA (fragmentos de DNA, cDNA ou oligonucleotídeos) fixados a um suporte sólido num padrão ordenado. Esses fragmentos de DNA (as sondas, *probes*) em geral correspondem a genes conhecidos. Os *microarrays* são baseados na hibridização de ácidos nucleicos, no qual as sondas são usadas para encontrar as sequências complementares. É aplicada uma solução contendo uma mistura de dois tipos de células de DNA ou RNA ao suporte sólido. Os ácidos nucleicos, o mRNA, DNA ou cDNA isolado de células experimentais na mistura, são marcados com um marcador fluorescente e aplicado no *array*. Qualquer uma das moléculas de DNA ou RNA que são complementares às sondas no *array* irá hibridizar com elas e emitir fluorescência, que pode ser detectada por um *scanner* apropriado. Os *microarrays* permitem a detecção de alelos específicos, polimorfismos de um único nucleótido (SNPs, pronuncia-se *snips*), e até mesmo proteínas particulares (Pierce [44]). Na Figura 1.3 esquematiza-se o processo experimental dos *microarrays* desde a hibridização até a obtenção dos dados para análise estatística.

A empresa Affymetrix, Inc., Santa Clara, California, USA, desenvolveu modelos para análise de muitos genes e também tem disponível software para análise da expressão de genes (ver `www.affimetrix.com`). Esta tecnologia está disponível em Portugal através da companhia *Genómica/ STAB VIDA*, Oeiras.

No Centro de Biologia Celular, do departamento de Biologia da Universidade de Aveiro (UA), estão a ser aplicadas técnicas de genómica (bioinformática, *arrays* de DNA, electroforese bidimensional e espectrometria de massa) com objectivo de identificar genes e os seus produtos cuja expressão é afectada pela descodificação ambígua do mRNA nalgumas espécies

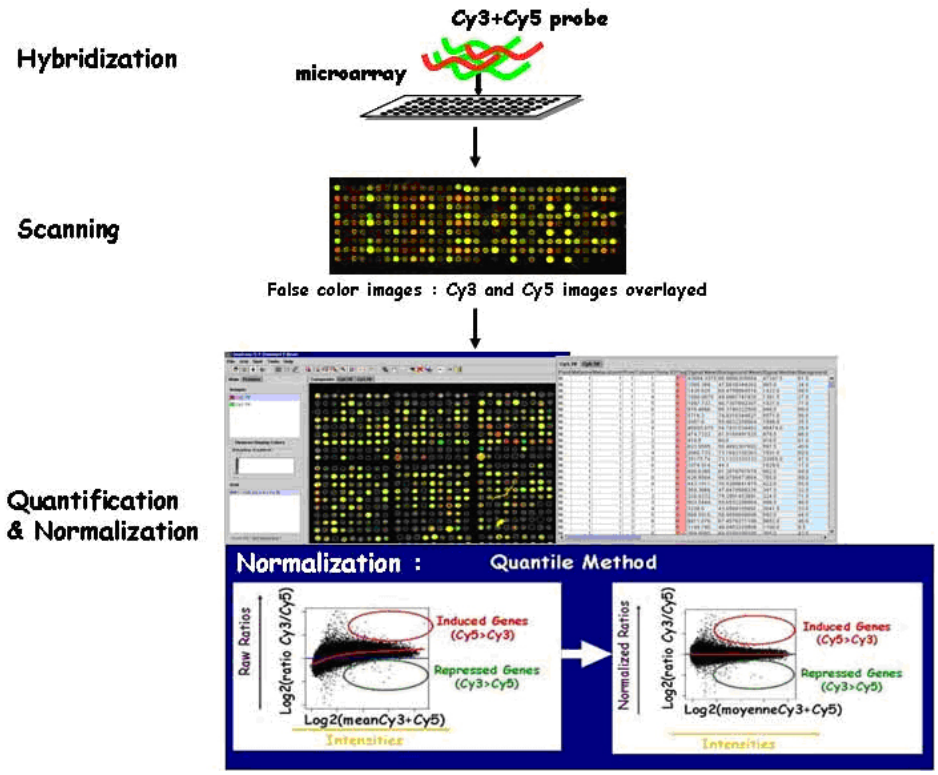


Figura 1.3: Processo de análise de *microarrays* (ima [9]). Para um chip de DNA, o mRNA de células experimentais e de controlo são marcados com nucleotídeos de fluorescência vermelha e verde, respectivamente (ou vice versa). Os cDNA marcados são misturados e hibridizados ao chip que contém sondas de genes diferentes. A hibridização dos cDNA vermelho e verde é proporcional às quantidades relativas de mRNA nas amostras. A fluorescência para cada ponto do *microarray* é avaliada com o microscópio de varredura e aparece uma só cor. O vermelho (verde) indica a hiperexpressão de um gene nas células marcadas a vermelho (verde). O amarelo indica igual expressão das células de controle e experimentais (hibridização igual de cDNAs vermelho e verde), e nenhuma cor indica a falta de expressão tanto nas células experimentais como nas de controle.

---

de leveduras. Desenvolveram também um sistema bioinformático chamado Anaconda (disponível em `http://bioinformatics.ua.pt/aplications/anaconda`) e modelos matemáticos para análise do contexto dos codões à escala genómica, em colaboração com investigadores dos Departamentos de Electrónica e de Matemática da UA (Santos et al. [47]).

Extraída a sequência genómica de qualquer espécie, o Anaconda permite obter uma tabela 61×64 (ver Tabela 1.1) dos resíduos ajustados de Pearson (denotados por d_{ij} em Pinheiro et al. [46]) resultante do teste χ^2 para a independência entre pares de codões consecutivos. Esta tabela permite, de algum modo, avaliar o grau de preferência dos pares consecutivos no

genoma.

	AAA ⋯ UAA UAG UGA UUU
AAA ⋮ UUU	resíduos ajustados de Pearson

Tabela 1.1: Tabela resultante do Anaconda.

1.2 Estatística Multivariada na Biologia Molecular

Uma análise estatística diz-se multivariada quando os dados incluem medições simultâneas de mais que uma variável ou característica. Estas podem ser quantitativas (discretas ou contínuas) ou qualitativas (ordenadas ou não em categorias). A análise estatística multivariada incorpora várias técnicas de análise de dados que procuram descrever, classificar e sumariar informação dos dados. Estas técnicas podem ser probabilísticas, inferenciais ou descritivas e é, sempre que possível, útil recorrer à utilização de software estatístico (R, SPSS, etc). Os métodos multivariados e os respectivos procedimentos variam consoante os objectivos da investigação. Genericamente, podemos classificar os diferentes objectivos de investigação do seguinte modo (Johnson and Wichern [33]):

- Redução dos dados ou simplificação estrutural.
- Agrupamento.
- Dependência entre as variáveis.
- Predição.
- Construção de hipóteses e testes.

Existem inúmeros estudos e artigos publicados onde a Estatística Multivariada é aplicada à Biologia Molecular. Nesta secção não pretendemos apresentar exaustivamente todos os tipos de análises multivariadas em dados de Biologia Molecular mas sim dar uma ideia de alguns tópicos e variedade de trabalhos que têm sido realizados.

No artigo de Idalino et al. [32], o objectivo do trabalho é conhecer as características associadas à similaridade e ao grau de polimorfismo presente em três espécies de animais (*Bois taurus*, *Ovis ares* e *Capra irca*) com base no gene HS70.1. Este gene provoca a mastite, doença muito comum na pecuária leiteira, responsável pela inflamação da glândula mamária e diminuição da produção de leite, comprometendo a sua qualidade a qual, por vezes, pode

levar à morte do animal. Através da Estatística Multivariada, nomeadamente do estudo de similaridade, compararam o alinhamento das sequências genéticas nas três espécies em estudo, bem como quantificaram os polimorfismos existentes.

Da análise de *microarrays* resulta um elevado número de genes e as técnicas de Estatística Multivariada para a redução da dimensionalidade são fulcrais. Por exemplo, o artigo de Gardner-Lubbe et al. [26], mostra as vantagens do uso dos biplots com dados de *microarrays* para visualizar as observações e os genes num único gráfico. Uma das técnicas usadas foi a Análise de Componentes Principais para redução da dimensionalidade. Aplicaram os métodos a uma base de dados sobre um grupo incubador e conseguiram a separação das classes.

No artigo de Landgrebe et al. [34], procuraram ver quais eram os efeitos de antidepressivos em ratos através da análise da diferença de expressão genética entre dois tratamentos. Os resultados foram obtidos com o método da Análise de Componentes Principais e concluíram que os perfis de expressão dos dois medicamentos são significativamente diferentes, sendo que quanto maior for a duração do tratamento maior será a diferença entre os perfis. Também concluíram que os medicamentos testados podem complementar-se embora actuem em vias diferentes.

No artigo de Ghosh and Chinnaiyan [27], desenvolveram novos algoritmos construídos com base em alguns já existentes. Estes estão baseados em mistura de modelos estatísticos para conglomerados. Analisaram dados de *microarrays* referentes à melanoma e ao cancro da próstata. Os resultados que obtiveram foram considerados muito satisfatórios pois em ambos os casos conseguiram a separação das classes. No entanto sugeriram que os algoritmos poderiam ser melhorados.

A Estatística Multivariada é também muito aplicada na área da Agricultura para melhorar o genótipo e fenótipo das espécies. Por exemplo, no artigo de Freitas et al. [24], avaliaram o desempenho de trinta populações de azevém (espécie de trigo) sob três regimes de água usando quatro variáveis. Para isso, utilizaram uma técnica da Estatística Multivariada com base na distância generalizada de Mahalanobis para estimar a dissemelhança genética. Puderam assim encontrar quais os cruzamentos dos genótipos que são indicados para melhorar os genótipos das gerações seguintes.

No artigo de Fonseca and Silva [23], procuraram estudar a divergência genética e identificar possível duplicidade em acessos de feijão, utilizando variáveis canónicas para os descritores agronómicos e fenológicos. Todas estas mostraram-se ser importantes para a descrição do germoplasma. No entanto, após a aplicação da Análise de Conglomerados, nomeadamente com o método do vizinho mais próximo, identificaram acessos repetidos de feijão em dois grupos. Recorrer à Estatística Multivariada foi muito útil porque, como referiram aqueles autores, além de económica (apenas exigiu cálculos), conseguiram identificar quais eram as

repetições de acesso e seleccionar os descritores utilizados nas actividades de caracterização.

No artigo de Martel et al. [37], o estudo consistia em aplicar técnicas de Estatística Multivariada para tentar discriminar as três raças e populações de palmeiras existentes ao longo da bacia dos rios Amazonas e Solimões. Nesse trabalho é salientado o carácter sócio-político do estudo, transcreve-se: "A pupunha tem um potencial económico e social muito grande como fonte de alimento para o homem e animais, sendo sem dúvida a palmeira mais importante da América pré-colombiana." A Análise de Componentes Principais, Análise Discriminante e Análise de Conglomerados forneceram bons resultados na discriminação das raças, para além de mostraram quais eram os descritores (variáveis) mais importantes.

1.3 Motivação e Organização da Dissertação

Um biplot é um gráfico que permite a visualização de dados multivariados num espaço reduzido. Com este tipo de representação gráfica podem ser detectados relacionamentos entre variáveis e/ou existência de grupos de indivíduos. A sua aplicação a dados reais de Biologia Molecular tem sido considerada nos últimos anos. No artigo de Park et al. [43] várias metodologias de biplots são estudadas e aplicadas a dados de *microarrays* relativos a 4 tipos de cancro. Não foram encontrados trabalhos publicados usando biplots no estudo do sequenciamento de codões. Motivados por este facto, e tendo como base o trabalho de Park et al. [43], na presente dissertação abordamos 3 técnicas de redução de dados e a sua metodologia correspondente nos biplots e aplicamo-las a dois conjuntos de dados reais da Biologia Molecular:

1. um conjunto já estudado no referido artigo, e
2. outro novo, relativo ao grau de preferência de pares de codões em sequências de DNA de 123 espécies cujo genoma foi extraído do ensembl, e processado com auxílio do Anaconda.

Esta dissertação é constituída, para além desta introdução, por mais três capítulos e dois apêndices.

No Capítulo 2 começamos por introduzir o conceito biplot, apresentamos um breve resumo do estado da arte e abordamos conceitos e metodologias de especial importância para o estudo dos biplots. De seguida, desenvolvemos algum material teórico sobre técnicas de redução da dimensionalidade dos dados: Análise de Componentes Principais, Análise de Correspondência Simples e Escalonamento Multidimensional. Finalmente, resumimos os procedimentos para a obtenção de biplots associados a cada uma daquelas três técnicas de redução e ilustramos com um exemplo de aplicação no software R. A base de dados que escolhemos para exemplificar

é o conjunto de dados conhecida por *Iris* e foi retirada do repositório `http://archive.ics.uci.edu/ml/datasets/Iris`. Trata-se de um conjunto de dados divididos em 3 classes de 50 indivíduos cada, onde cada classe representa uma espécie: *Iris Setosa*, *Iris Versicolor* e *Iris Virginica*. Este conjunto de dados é caracterizado por quatro variáveis que exprimem o comprimento e largura da sépala e o comprimento e a largura da pétala. Salientamos que os biplots foram obtidos quer directamente, usando comandos específicos do R, quer usando uma sequência de comandos aqui proposta com o objectivo de construir biplots com maior flexibilização nas etiquetas e coloração dos pontos.

No Capítulo 3 aplicamos todas as técnicas abordadas no Capítulo 2 a dois conjuntos de dados reais. Concretamente, analisamos os seguintes conjuntos de dados:

- Dados 1: dados de nível de expressão genética relativos ao cancro do colón, retirados de Alon [11]. Estes dados são referentes ao nível de expressão de dois mil genes medidos através da tecnologia de *microarrays* em sessenta e dois tecidos;

- Dados 2: dados de grau de preferência de pares de codões consecutivos em sequências completas de DNA de 123 espécies. As sequências de cDNA (*abinitio prediction transcript*) foram extraídas do ensembl (`www.ensembl.org`) em Janeiro 2010, e processadas usando o software Anaconda (`http://bioinformatics.ua.pt/aplications/anaconda`).

Com os Dados 1 comparamos os nossos resultados com os obtidos em (Park et al. [43]). Com os Dados 2 tentamos extrair informação a nível de possíveis agrupamentos de reinos.

No Capítulo 4, apresentamos um breve resumo das conclusões obtidas no estudo desenvolvido no capítulo anterior, e ainda mencionamos possíveis trabalhos que poderão ser desenvolvidos posteriormente.

Os apêndices estão divididos em duas secções. Na primeira consta todo o código realizado nos estudos presentes ao longo desta dissertação. Na segunda está representada a tabela referente às espécies estudadas para o conjunto Dados 2.

Capítulo 2

Métodos Biplot

2.1 Conceitos Fundamentais

O conceito de biplot deve-se ao trabalho de Gabriel [25], sendo a ideia fundamental construir uma representação simultânea de indivíduos/observações e variáveis num espaço de dimensão 2 ou 3.

Os biplots podem considerar-se como uma generalização dos gráficos de dispersão. Os diagramas de dispersão permitem representar até três variáveis, sendo que cada um dos eixos representa uma variável e as etiquetas dos pontos observados podem representar uma terceira variável. O biplot generaliza a seguinte ideia, representar um gráfico de dispersão contendo marcadores para as variáveis e marcadores para os indivíduos observados (Vairinhos and Lobo [49]).

Genericamente, seja $\mathbf{X}_{n\times p}$ uma matriz de dados tal que:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \tag{2.1}$$

onde o elemento x_{ij} representa o valor da j-ésima variável no i-ésimo individuo, sendo

$$\begin{aligned} n &= \text{número de indivíduo em estudo,} \\ p &= \text{número de medições realizadas em cada indivíduo.} \end{aligned}$$

O método biplot gera uma representação procurando que:

- as distâncias entre os indivíduos projectados num espaço de dimensão reduzida sejam próximas às originais;

- a projecção dos indivíduos nos eixos sejam as mais próximas das originais;
- os vectores que representam as variáveis originais ajudem a verificar quanto de peso cada novo eixo dá a cada variável original;
- o cosseno do ângulo entre os vectores que representam as variáveis originais se aproxime da correlação entre essas variáveis.

Para estes objectivos, torna-se imperativo decompor a matriz dos dados.

Decomposição em Valores Singulares

A decomposição em valores singulares de uma matriz genérica é um dos resultados mais importantes na Teoria de Matrizes, pois permite factorizar qualquer matriz, mesmo sendo rectangular, de forma simultaneamente simples e poderosa (Cadima [15]). É com base neste resultado que Gabriel desenvolveu a teoria dos biplots (Gabriel [25]).

Qualquer matriz de dados $\mathbf{X}_{n\times p}$ definida por (2.1), com característica k ($k \leq \min(n,p)$) e com r ($r \leq k$) valores singulares não nulos, $\sigma_1, \cdots, \sigma_r$, é factorizável na forma

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{2.2}$$

chamada decomposição em valores singulares (*Singular Value Decomposition* - SVD) da matriz $\mathbf{X}$, sendo $\mathbf{V}_{p\times p} = [\ \mathbf{v}_1 \ \cdots \ \mathbf{v}_p\ ]$ uma matriz ortogonal construída a partir de um conjunto ortonormado de vectores próprios $\mathbf{v}_i$ da matriz $\mathbf{U}_{n\times n} = [\ \mathbf{u}_1 \ \cdots \ \mathbf{u}_n\ ]$ uma matriz cujas colunas são determinadas por $\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{X}\mathbf{v}_i$, e $\mathbf{\Sigma}_{n\times p}$ é uma matriz definida da seguinte forma

$$\mathbf{\Sigma} = \begin{pmatrix} & & & 0 & \dots & 0 \\ & \mathbf{Diag} & & \vdots & \ddots & \vdots \\ & & & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix}, \text{onde} \quad \mathbf{Diag} = \begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_r \end{pmatrix} \tag{2.3}$$

As colunas $\mathbf{u}_i$ de $\mathbf{U}$ são os vectores singulares à esquerda, e $\mathbf{v}_i$, as colunas de $\mathbf{V}$, são os vectores singulares à direita. Analogamente à decomposição espectral de uma matriz simétrica, a matriz $\mathbf{X}_{n\times p}$, pode ser escrita da seguinte forma

$$\mathbf{X} = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^T = \sum_{i=1}^{r}\sigma_i\mathbf{u}_i\mathbf{v}_i^T. \tag{2.4}$$

Observações 2.1.1. *1. Os valores singulares σ_i estão ordenados por ordem decrescente.*

2. *A decomposição de uma matriz é sempre possível, mas apenas é única se não houver valores singulares repetidos. Esta questão resulta directamente da discussão sobre a existência e unicidade da decomposição espectral das matrizes* $\boldsymbol{X}\boldsymbol{X}^T$ *e* $\boldsymbol{X}^T\boldsymbol{X}$ *(Cadima [15]).*

3. *Se* $\boldsymbol{X}$ *tem decomposição em valores singulares dada pela equação (2.2), então a transposta de* $\boldsymbol{X}$ *tem a decomposição em valores singulares dada por* $\boldsymbol{X}^T = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T$.

Teorema 2.1.1. *Seja* $\boldsymbol{X}_{n\times p}$ *uma matriz de característica k. Dada a decomposição em valores singulares* $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, *a matriz* $\boldsymbol{Y}_{n\times p}$ *de característica* $m < k$ *que melhor aproxima* $\boldsymbol{X}$, *no sentido de minimizar a distância matricial* $\|\boldsymbol{X} - \boldsymbol{Y}\| = \sqrt{\sum_i \sum_j (x_{ij} - y_{ij})^2}$, *é dada por:*

$$\boldsymbol{Y} = \boldsymbol{U}_m \boldsymbol{\Sigma}_m \boldsymbol{V}_m^T$$

onde, $\boldsymbol{U}_m$ *e* $\boldsymbol{V}_m$ *são as matrizes constituídas pelas m colunas de* $\boldsymbol{U}$ *e* $\boldsymbol{V}$, *respectivamente, associadas aos m maiores valores singulares e* $\boldsymbol{\Sigma}_m$ *a matriz diagonal* $m \times m$ *resultante de reter apenas as linhas e colunas de* $\boldsymbol{\Sigma}$ *associadas aos m maiores valores singulares.*

Observações 2.1.2. 1. $\boldsymbol{Y} = \boldsymbol{U}_m \boldsymbol{\Sigma}_m \boldsymbol{V}_m^T$ *é uma decomposição em valores singulares de* $\boldsymbol{Y}$.

2. *Usando a forma* $\boldsymbol{X} = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$, *da decomposição em valores singulares de* $\boldsymbol{X}$, *verifica-se que* $\boldsymbol{Y}$ *é a matriz que se obtem retendo apenas as m primeiras parcelas da decomposição em valores singulares de* $\boldsymbol{X}$.

Biplots

Vamos então considerar a decomposição em valores singulares de uma matriz centrada de dados e apresentar de maneira simplificada a essência dos biplots.

Seja a matriz de dados $\mathbf{X}_{n\times p}$ com característica k, factorizável por (2.2), ou seja, $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$. Para $0 \leq \alpha \leq 1$, a decomposição em valores singulares pode ser escrita da forma seguinte:

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}^{\alpha}\boldsymbol{\Sigma}^{1-\alpha}\mathbf{V}^T \tag{2.5}$$

Definindo:

$$\mathbf{G} = \mathbf{U}\boldsymbol{\Sigma}^{\alpha} \tag{2.6}$$

$$\mathbf{H} = \mathbf{V}\boldsymbol{\Sigma}^{1-\alpha} \tag{2.7}$$

obtemos:

$$\mathbf{X} = \mathbf{G}\mathbf{H}^T \tag{2.8}$$

Como a matriz $\mathbf{G}$ tem dimensão $n \times k$ existe uma correspondência entre as linhas de $\mathbf{G}$ e os indivíduos da matriz $\mathbf{X}$. A matriz $\mathbf{H}$ tem dimensão $p \times k$, existe uma correspondência entre as linhas de $\mathbf{H}$ e as variáveis da matriz $\mathbf{X}$.

A partir da equação (2.8) podemos dizer que o elemento (i, j) da matriz centrada de dados $\mathbf{X}$ é obtido através do produto interno da i-ésima linha de $\mathbf{G}$ com a j-ésima coluna de $\mathbf{H}^T$. Portanto:

$$x_{ij} = \langle \mathbf{g}_i, \mathbf{h}_j \rangle = \mathbf{g}_i^T \mathbf{h}_j \tag{2.9}$$

onde $\mathbf{g}_i^T$ representa a i-ésima linha de $\mathbf{G}$ e $\mathbf{h}_j$ representa a j-ésima linha de $\mathbf{H}$. Ambos os vectores $\mathbf{g}_i$ e $\mathbf{h}_j$ são vectores k-dimensional. Segue-se então que no espaço $\mathbb{R}^k$ é possível representar cada indivíduo pela linha da matriz $\mathbf{G}$ que lhe está associada, e cada variável pela linha da matriz $\mathbf{H}$ correspondentes, de modo que o produto interno entre estes seja igual ao elemento x_{ij} da matriz $\mathbf{X}$.

Quando aplicado na prática, nem sempre é possível obter uma representação gráfica exacta, a não ser que a matriz $\mathbf{X}$ tenha característica $k = 2$ ou $k = 3$. Porém, pode-se representar graficamente em $\mathbb{R}^2$ ou $\mathbb{R}^3$ os indivíduos com marcadores cujas coordenadas são apenas os dois ou três primeiros elementos de cada vector $\mathbf{g}_i$, e no mesmo sistema de eixos, representar as variáveis com coordenadas dados pelos primeiros elementos de $\mathbf{h}_j$. A esta representação chamamos de **biplot**. Chamamos de marcadores dos indivíduos no biplot aos subvectores $\mathbf{g}_i^{(r)}$ que são constituídos pelas r primeiras coordenadas dos vectores $\mathbf{g}_i$. Chamamos de marcadores das variáveis no biplot aos subvectores $\mathbf{h}_j^{(r)}$ que são constituídos pelas r primeiras coordenadas dos vectores $\mathbf{h}_j$ (Cadima [15]).

Consoante o valor que α toma em (2.6) e (2.7) obtém-se diferentes biplots.

Biplot com $\alpha = 0$

Tomando $\alpha = 0$ as matrizes $\mathbf{G}$ e $\mathbf{H}$ definidas em (2.6) e (2.7) ficam do seguinte modo:

$$\mathbf{G} = \mathbf{U} \quad \text{e} \quad \mathbf{H} = \mathbf{V}\mathbf{\Sigma} \tag{2.10}$$

Deste modo, a factorização (2.2) preserva a métrica das colunas (*Column Metric Preserving*). O biplot que lhe é associado é habitualmente designado na literatura por GH biplot e corresponde ao biplot clássico de Gabriel. Este biplot satisfaz as seguintes propriedades:

- A projeccção ortogonal dos marcadores de indivíduos sobre o subespaço gerado pelo marcador da variável j é proporcional aos valores dos indivíduos na variável j.
- O cosseno do ângulo entre cada par de linhas da matriz $\mathbf{H}$ é o coeficiente de correlação entre as respectivas variáveis.

- O produto interno entre cada par de linhas da matriz $\mathbf{H}$ é proporcional à covariância entre as respectivas variáveis.
- A norma de cada linha da matriz $\mathbf{H}$ é proporcional ao desvio padrão da respectiva variável.
- A distância euclidiana entre cada par de linhas da matriz $\mathbf{G}$ é proporcional à distância de Mahalanobis entre os respectivos indivíduos.

Biplot com $\alpha = 1$

Tomando $\alpha = 1$, as matrizes $\mathbf{G}$ e $\mathbf{H}$ definidas em (2.6) e (2.7) ficam do seguinte modo:

$$\mathbf{G} = \mathbf{U\Sigma} \quad \text{e} \quad \mathbf{H} = \mathbf{V} \tag{2.11}$$

Deste modo a factorização (2.2) preserva a métrica das linhas (*Row Metric Preserving*). O biplot que lhe é associado é habitualmente designado na literatura por JK biplot. Este biplot satisfaz as seguintes propriedades:

- a nuvem de n pontos no subespaço principal tem dimensão r (habitualmente, selecciona-se $r = 2$ ou 3 para representar os r primeiros elementos dos vectores $\mathbf{g}_i$ e $\mathbf{h}_j$);
- as distâncias euclidianas entre os marcadores de indivíduos serão, aproximadamente, as distâncias euclidianas entre os indivíduos no espaço $\mathbb{R}^p$;
- o produto interno entre os marcadores de indivíduos e de variáveis dará uma aproximação de x_{ij}.

Biplot com $\alpha = 0.5$

Tomando $\alpha = 0.5$, as matrizes $\mathbf{G}$ e $\mathbf{H}$ definidas em (2.6) e (2.7) ficam do seguinte modo:

$$\mathbf{G} = \mathbf{U\Sigma}^{0.5} \quad \text{e} \quad \mathbf{H} = \mathbf{V\Sigma}^{0.5} \tag{2.12}$$

O biplot que lhe é associado é habitualmente designado na literatura por SQRT biplot. Com este tipo de representação atinge-se a mesma qualidade de representação para os indivíduos e variáveis mas não a máxima que é possível separadamente para os indivíduos e variáveis.

Resumindo, se queremos obter uma maior qualidade de representação das variáveis, de acordo com a decomposição (2.5) usamos $\alpha = 0$. Para obter uma maior qualidade de representação dos indivíduos usamos $\alpha = 1$, se $\alpha = 0.5$ estamos em situação de compromisso. De acordo com Galindo (Villardón [53]), e embora Gabriel tenha afirmado que o nome do método

biplot é devido à técnica permitir a representação conjunta de todas as linhas da matriz, esta representação não é simultânea no sentido estrito uma vez que a qualidade de representação não é a mesma para as linhas e para as colunas.

HJ-Biplot

Galindo (Villardón [53]), criou um novo método chamado HJ-biplot que permite uma representação simultânea dos indivíduos e variáveis no sentido estrito porque preserva a métrica das linhas e a métrica das colunas (*Row Column Metric Preserving*). Os marcadores das variáveis são definidos por:

$$\mathbf{H} = \mathbf{U\Sigma}$$

e os marcadores dos indivíduos são definidos por:

$$\mathbf{J} = \mathbf{V\Sigma}$$

Mas tendo em conta que

$$\mathbf{X} = \mathbf{U\Sigma V} \neq \mathbf{HJ}^T = \mathbf{U\Sigma}^2\mathbf{V}$$

ou seja, não se garante a decomposição dos valores observados como produtos internos de marcadores, mas garante-se a qualidade de representação simultânea para as variáveis e indivíduos (Vairinhos and Galindo [48]). Galindo e Cuadras (1986) propuseram algumas medidas para melhor interpretar os gráficos resultantes de uma análise HJ-Biplot (Dourado et al. [19]).

O método HJ-Biplot tem então as seguintes propriedades (Alonso [12]) :

- Esta representação fornece a melhor interpretação simultânea.
- Os produtos escalares das colunas da matriz $\mathbf{X}$ coincidem com os produtos escalares dos marcadores $\mathbf{H}$.
- O cosseno do ângulo entre dois vectores $h_i h_j$ representa a correlação entre as variáveis x_i e x_j.
- Os produtos escalares das linhas da matriz $\mathbf{X}$ coincidem com os produtos escalares dos marcadores $\mathbf{J}$.
- A distância euclidiana entre as linhas da matriz $\mathbf{X}$ coincide com a distância euclidiana entre os marcadores $\mathbf{J}$.
- Os marcadores para as linhas coincidem com as coordenadas dos indivíduos dentro do espaço das CPs das variáveis.

- Os marcadores para as colunas coincidem com as coordenadas das variáveis no espaço das componentes das linhas.

- Se uma variável tem um valor que conduz a um indivíduo, o ponto que representa essa variável estará próximo do ponto que representa o individuo.

- A proximidade entre os indivíduos é interpretada em termos de semelhanças.

- Quanto maior as distâncias que aparecem nos pontos representativos dos marcadores de coluna ao centro de gravidade (centróide), maior será a variabilidade no estudo.

- Quanto menor o ângulo entre os vectores que ligam os pontos que representam duas variáveis com o centro de gravidade, mais correlacionadas estarão as variáveis.

- A qualidade de representação para as linhas e colunas é a mesma e vem expressada por: $\left(\frac{\sum_{i=1}^{d}\lambda_i^2}{\sum_{i=1}^{r}\lambda_i^2}\right) \times 100$.

Em 1996 foi publicado o primeiro livro sobre a teoria dos biplots. Os seus autores, Gower e Hand aplicaram a teoria dos biplots na Análise de Componentes Principais (ACP), desenvolveram a teoria dos biplots lineares para Escalonamento Multidimensional (Multidimensional Scaling - MDS), previsão e interpolação dos dados e eixos. Aplicaram-na igualmente na Análise Múltipla de Correspondência, falaram ainda sobre os biplots canónicos, biplots não lineares, e fizeram uma generalização sobre a teoria dos biplots.

No artigo de Blasius et al. [14], os seus autores mostraram formas úteis de exibir biplots e desenvolveram um programa para sua construção automática. Chamaram à atenção para os possíveis defeitos dos biplots:

1. podem ser deselegantes, muitas das vezes são apresentadas as escalas horizontal e vertical mesmo quando não é necessário;

2. têm escala desigual nas direcções x e y;

3. inconveniência de avaliação dos produtos internos;

4. o modo como é feita a etiquetagem das variáveis e dos indivíduos pode interferir com a interpretação se as etiquetas ficarem sobrepostas com os pontos da amostragem ou das variáveis, respectivamente.

2.2 Redução da dimensionalidade dos dados

2.2.1 Análise de Componentes Principais

A Análise de Componentes Principais (ACP) permite reduzir o conjunto original de variáveis num menor conjunto de variáveis não correlacionadas (componentes principais, CPs) de modo a representar a maior informação, explicando a máxima variabilidade do conjunto original das variáveis. Espera-se que a representação dos indivíduos no subespaço reduzido definido pelas primeiras CPs permita descobrir relações e propriedades entre eles.

Dada uma matriz $\mathbf{X}_{n\times p}$, que contém de n indivíduos medidos em $p = 2$ variáveis centradas, os n indivíduos podem ser representados no espaço $\mathbb{R}^2$ por n pontos P_i de coordenadas (x_1, x_2) para OX_1 e OX_2 e, ao mesmo tempo (y_1, y_2) para OY_1 e OY_2 (Figura 2.1). Se a transformação

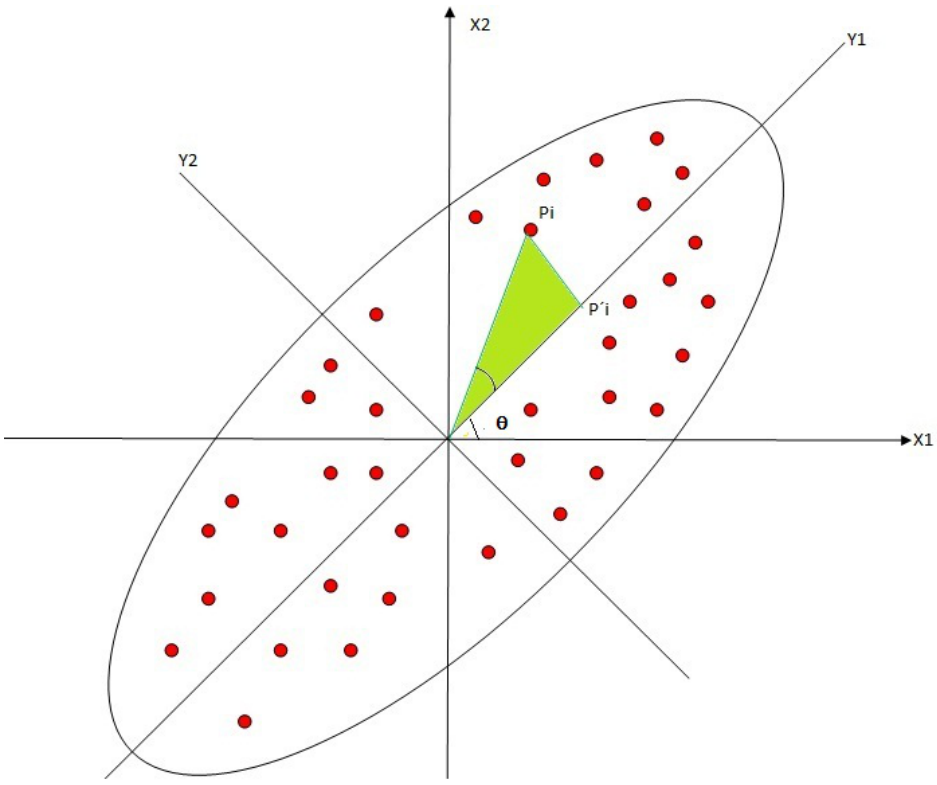


Figura 2.1: Projecção dos pontos na 1ª CP (adaptada de Villardón [52] e Álvarez González [35]).

de eixos ocorre através de uma rotação θ, então

$$\begin{aligned} y_1 &= x_1 \cos\theta + x_2 \sin\theta \\ y_2 &= -x_1 \sin\theta + x_2 \cos\theta \end{aligned}$$

Segundo a Figura 2.1, para obter a recta que representa a maior dispersão, projectando os indivíduos nela própria, deve-se determinar o ângulo de rotação θ que satisfaça esta condição. Tal condição consiste em minimizar $\sum_{i=1}^{n}(P_iP_i^{'})^2$. Como:

$$(OP_i)^2 = (OP_i^{'}) + (PP_i^{'})^2$$

será necessário minimizar $\left[\frac{1}{n-1}\sum_{i=1}^{n}(PP_i^{'})^2\right]$. Tal equivale a maximizar $\left[\frac{1}{n-1}\sum_{i=1}^{n}(OP_i^{'})^2\right]$. Uma vez obtida a direcção do eixo OY_1, OY_2 será perpendicular a OY_1. Se tivermos um conjunto de dimensão $p \geq 2$ aplicamos sucessivamente a ideia acima descrita, que nos permitirá obter as direcções principais.

Observações 2.2.1.

- *A técnica ACP não requere uma distribuição normal multivariada para os dados.*
- *No caso dos dados provirem de uma população normal multivariada, as CPs podem ser interpretadas em termos de elipsóides de densidade constantes.*

Sejam:

(a) $\mathbf{X} = [X_1 \quad X_2 \quad \cdots \quad X_p]$ um vector aleatório de p componentes[1].

(b) $\mathbf{S}$ = matriz de covariâncias de $\mathbf{X}$ com valores próprios $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

(c) $\mathbf{l}_i = [l_{1i} \quad l_{2i} \quad \cdots \quad l_{pi}]$, $i = 1, 2, \cdots, p$, p vectores de constantes reais.

A ACP procura uma nova combinação linear das variáveis originais:

$$\begin{aligned} Y_1 &= \mathbf{l}_1\mathbf{X}^T = l_{11}X_1 + l_{21}X_2 + \cdots + l_{p1}X_p \\ Y_2 &= \mathbf{l}_2\mathbf{X}^T = l_{12}X_1 + l_{22}X_2 + \cdots + l_{p2}X_p \\ &\vdots \\ Y_p &= \mathbf{l}_p\mathbf{X}^T = l_{1p}X_1 + l_{2p}X_2 + \cdots + l_{pp}X_p \end{aligned} \tag{2.13}$$

de modo que os vectores de constantes $\mathbf{l}_i$, $i = 1, 2, \cdots, p$, tenham norma unitária e que as novas variáveis aleatórias $Y_1, Y_2, \cdots, Y_p$ sejam não correlacionadas e com variância máxima.

Proposição 2.2.1. *Nas condições anteriores, sendo $\boldsymbol{S}_{p\times p}$ a matriz de variâncias-covariâncias de $\boldsymbol{X}_{n\times p}$, a variância de $\boldsymbol{Y}$ é igual a $Var(\boldsymbol{Y}) = \boldsymbol{lSl}^T$.*

Define-se a primeira CP $Y_1 = \mathbf{l}_1\mathbf{X}^T$ como combinação linear das variáveis originais que maximizam $\mathbf{l}_1\mathbf{S}\mathbf{l}_1^T$, sujeito a $\mathbf{l}_1\mathbf{l}_1^T = 1$.

[1] Assume-se que $p \leq n$. Quando $n < p$, as CPs são obtidas recorrendo à SVD.

Proposição 2.2.2. *Nas condições anteriores* $\boldsymbol{l}_1$ *é o vector próprio unitário correspondente ao maior valor próprio* λ_1 *de* $\boldsymbol{S}$*. Além disso,* $Var(Y_1) = \boldsymbol{l}_1 \boldsymbol{S}\ \boldsymbol{l}_1^T = \lambda_1$.

Continuando o processo, encontramos todas as combinações lineares $Y_k = \mathbf{l}_k \mathbf{X}^T$ (até p) com as seguintes condições:

- maximizar $\mathbf{l}_k \mathbf{S} \mathbf{l}_k^T$,
- $\mathbf{l}_k \mathbf{l}_k^T = 1$,
- $\mathbf{l}_k \mathbf{l}_i^T = 0 \quad (i \neq k)$.

Por outras palavras, as CPs são então dadas pelos valores próprios e vectores próprios ortonormais de $\mathbf{S}$ de modo que os valores próprios estejam ordenados por ordem decrescente.

Proposição 2.2.3. *Nas condições anteriores, as p CPs são dadas por:*

$$Y_k = \boldsymbol{l}_k \boldsymbol{X}^T, (k = 1, \cdots, p)$$

com $Var(Y_k) = \lambda_k$ *e* $Cov(Y_i, Y_k) = 0$, $i \neq k$.

Demonstração. Ver página 342 de (Johnson and Wichern [33]). □

Proposição 2.2.4. *Considerando as condições anteriores, a variância total verifica*

$$\sum_{i=1}^{p} Var(Y_i) = \lambda_1 + \cdots + \lambda_p = \textit{traço } (\boldsymbol{S})$$

Demonstração. Ver página 343 de (Johnson and Wichern [33]). □

Como consequência da Proposição 2.2.4 tem-se que a proporção de variância total devida à i-ésima componente principal é igual a:

$$\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_p}$$

Obtenção das Componentes Principais a partir da SVD

Seja $\mathbf{D}$, de elemento genérico d_{ij}, a matriz que resulta após centrarmos as variáveis, ou seja,

$$d_{ij} = x_{ij} - \overline{x}_j, \tag{2.14}$$

onde $\overline{x}_j = \dfrac{1}{n} \sum_{i=1}^{n} x_{ij}$. A aplicação da SVD à matriz proporciona o cálculo das coordenadas de representação das observações no subespaço das CPs, e a definição de marcadores para as linhas e para as colunas na construção do biplot da ACP simples.

Nas representações biplots, os indivíduos são normalmente identificados por pontos no subespaço reduzido, enquanto que as variáveis são representadas por vectores.

Tanto a utilização das CPs para analisar os indivíduos, como para reduzir a dimensão das variáveis, sugerem que estas sejam de máxima utilidade quando todas as variáveis medidas são pelo menos comparáveis em relação a sua variância, assim como as suas unidades de medidas. Quando isto não ocorre, muitos estatísticos sugerem estandardizar os dados. Neste caso, a matriz das covariâncias associada à estandardização de $\mathbf{X}$ passa a ser a matriz de correlações de $\mathbf{X}$ e o elemento genérico de (2.14) da matriz $\mathbf{D}$ é substituído por:

$$d_{ij} = (x_{ij} - \overline{x}_j)/s_j$$

onde s_j é o desvio padrão da variável j.

À matriz $\mathbf{D}$ que acabámos de obter aplica-se a SVD. A soma dos valores próprios, ou seja, o traço da matriz das correlações, irá coincidir com o número das p variáveis.

Uma outra variante é considerar a matriz $\mathbf{D}$ definida do modo seguinte:

$$d_{ij} = (x_{ij} - \overline{x}_j)/(s_j\sqrt{n-1}). \tag{2.15}$$

Isto é conseguido através de uma deformação da dispersão, restringindo a representação das variáveis no biplot a vectores de comprimento inferior ou igual à unidade. Portanto, todos os vectores das variáveis serão encontrados numa esfera de raio unitário. Neste caso, a soma dos valores próprios da matriz $\mathbf{D}$ definida por (2.15) será igual a $(n-1)p$.

Comparação entre a utilização da matriz das covariâncias e matriz das correlações

Observe-se que (Álvarez González [35]):

- A principal razão para utilizar a matriz de correlações em vez da matriz das covariâncias é que os resultados obtidos para as análises para diferentes conjuntos de variáveis são mais directamente comparáveis.
- Uma grande desvantagem da ACP baseada na matriz das covariâncias é a sensibilidade das CPs nas unidades de medida para cada variável. Se existe diferenças muito elevadas entre as variâncias, as variáveis mais dispersas tenderão a dominar as primeiras CPs.
- Uma propriedade das CPs baseadas na matriz das correlações é que, se em vez de normalizar $\mathbf{l}_k\mathbf{l}_k^T = 1$ se utilizar $\tilde{\mathbf{l}}_k\tilde{\mathbf{l}}_k^T = \lambda_k$ $(k = 1, \cdots, p)$ (λ_k é o valor próprio associado a $\mathbf{l}_k$), então $\tilde{l}_{kj}$, o j-ésimo elemento de $\tilde{\mathbf{l}}_k$, corresponde à correlação entre a variável j e a componente principal k.

2.2.2 Análise de Correspondências

As primeiras considerações matemáticas a respeito da Análise de Correspondências (AC) foram feitas por Hirschfeld (1935). A partir daí, os procedimentos numéricos e algébricos foram aplicados em diferentes contextos, nomeadamente em Ecologia e Psicologia. O método foi "redescoberto" na França no início da década de 60 e tem sido extensivamente usado naquele país como um método gráfico de análise de dados. A partir de 1975, a técnica tem sido utilizada em diversas áreas do conhecimento e em publicações em diversos idiomas (Greenacre and Hastie [30]).

A AC é uma técnica de análise exploratória de dados multivariados adequada para analisar tabelas de duas entradas ou tabelas de múltiplas entradas, tendo em conta algumas medidas de correspondência entre linhas e colunas. Basicamente, esta técnica converte uma matriz de dados não negativos num tipo particular de representação gráfica em que as linhas e colunas da matriz são simultaneamente representadas por pontos de um plano num gráfico. Este método permite estudar as relações e semelhanças existentes:

- entre as categorias em linhas e entre as categorias em colunas de uma tabela de contingência,

- entre o conjunto de categorias em linhas e o conjunto de categorias em colunas.

A AC mostra como as variáveis dispostas em linhas e colunas estão relacionadas e não somente se a relação existe. Embora seja considerada uma técnica descritiva e exploratória, a AC simplifica dados complexos e produz análises exaustivas de informações que suportam conclusões a respeito das mesmas. É altamente flexível quanto a pressuposições sobre os dados: o único requisito é ser aplicada a uma matriz rectangular com entradas não negativas. Observe-se que é possível transformar qualquer característica quantitativa em qualitativa, realizando-se uma partição de seu domínio de variação em classes.

A AC é mais efectiva se a matriz de dados for bastante grande, de modo a que a inspecção visual ou análise estatística simples não consegue revelar a sua estrutura. A AC pode ser considerada como um caso especial da ACP, porém dirigida a dados categóricos organizados em tabelas de contingência e não a dados contínuos. O objectivo do procedimento da AC é análogo a encontrar a maior componente principal de um conjunto de n indivíduos e p variáveis, com modificações devido à ponderação das observações e à métrica ponderada.

Teoria Básica

A forma mais simples de AC é a sua aplicação a uma tabela de contingência de dupla entrada sendo denominada Análise de Correspondência Simples. Seja $\mathbf{N}$ uma tabela de contingência de frequências absolutas e n o número total de observações associado à tabela. Sejam ainda, A e B dois factores de classificação com a níveis e b níveis, respectivamente. Suponhamos que o factor A está associado às linhas de $\mathbf{N}$ e o factor B às colunas de $\mathbf{N}$. Então $\mathbf{N}$ é representada do modo seguinte:

$$\mathbf{N} = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1j} & \cdots & n_{1,b-1} & n_{1,b} \\ n_{21} & n_{22} & \cdots & n_{2j} & \cdots & n_{2,b-1} & n_{2,b} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ n_{i1} & n_{i2} & \cdots & n_{ij} & \cdots & n_{i,b-1} & n_{i,b} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ n_{a-1,1} & n_{a-1,2} & \cdots & n_{a-1,j} & \cdots & n_{a-1,b-1} & n_{a-1,b} \\ n_{a,1} & n_{a,2} & \cdots & n_{a,j} & \cdots & n_{a,b-1} & n_{a,b} \end{bmatrix} \tag{2.16}$$

onde n_{ij} indica a frequência absoluta relativamente ao nível i do factor A com o nível j do factor B.

A matriz de frequências relativas será $\mathbf{T} = (1/n)\mathbf{N}$ e é chamada de matriz de correspondências. Cada linha ou coluna de $\mathbf{T}$ pode ser considerada um vector de proporções. Portanto, $\mathbf{T}$ devolve as frequências relativas de cada combinação de níveis dos factores A e B e representa-se do modo seguinte:

$$\mathbf{T} = \begin{bmatrix} \frac{n_{11}}{n} & \frac{n_{12}}{n} & \cdots & \frac{n_{1j}}{n} & \cdots & \frac{n_{1,b-1}}{n} & \frac{n_{1,b}}{n} \\ \frac{n_{21}}{n} & \frac{n_{22}}{n} & \cdots & \frac{n_{2j}}{n} & \cdots & \frac{n_{2,b-1}}{n} & \frac{n_{2,b}}{n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \frac{n_{i1}}{n} & \frac{n_{i2}}{n} & \cdots & \frac{n_{ij}}{n} & \cdots & \frac{n_{i,b-1}}{n} & \frac{n_{i,b}}{n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \frac{n_{a-1,1}}{n} & \frac{n_{a-1,2}}{n} & \cdots & \frac{n_{a-1,j}}{n} & \cdots & \frac{n_{a-1,b-1}}{n} & \frac{n_{a-1,b}}{n} \\ \frac{n_{a,1}}{n} & \frac{n_{a,2}}{n} & \cdots & \frac{n_{a,j}}{n} & \cdots & \frac{n_{a,b-1}}{n} & \frac{n_{a,b}}{n} \end{bmatrix} \tag{2.17}$$

A partir da matriz de correspondências definem-se os vectores $\mathbf{r}$ e $\mathbf{c}$ de frequências relativas marginais (em relação ao total geral n) denominados *massas*. Portanto, o vector $\mathbf{r}$ é o vector de massas de linhas constituído pelos a elementos que são as frequências relativas associadas a cada linha, ou seja, $\mathbf{r} = [r_1 \cdots r_a]$ onde a frequência relativa da linha i de $\mathbf{T}$ é:

$$r_i = \frac{n_{i+}}{n},$$

n_{i+} representa a frequência total observada na i-ésima categoria de A e $\mathbf{r}$. Analogamente, o vector $\mathbf{c}$ é vector de massas de colunas constituído pelos b elementos que são as frequências relativas associadas a cada coluna, ou seja, $\mathbf{c} = [c_1 \cdots c_b]$ onde a frequência relativa da coluna j de $\mathbf{T}$ é:

$$c_j = \frac{n_{+j}}{n}$$

n_{+j} representa a frequência total observada na j-ésima categoria de B e $\mathbf{c}$.

Nota 2.2.1. *Nos casos em que n_{i+} ou n_{+j} serem nulos, as categorias correspondentes devem ser retiradas da análise.*

Os vectores $\mathbf{r}$ e $\mathbf{c}$ são então estimativas das distribuições de probabilidades marginais, associadas aos factores A e B, respectivamente.

Designa-se por perfil da linha i o conjunto das frequências observadas para cada elemento dessa linha, relativamente ao total de observações nessa linha. Então, o perfil da linha i é dado pelos b valores, tal que:

$$pl_j^{(i)} = \frac{n_{ij}}{n_{i+}}, \qquad j = 1, \cdots, b$$

Necessitamos atribuir uma massa a cada ponto porque, ao introduzir os perfis, padronizamos os perfis, perdendo assim a informação sobre a proporção de indivíduos em cada linha (Pamplona [42]).

Matricialmente, define-se a matriz dos perfis de linha $\mathbf{P_L}$ por:

$$\mathbf{P_L} = \mathbf{Diag_r^{-1}T},$$

onde $\mathbf{Diag_r^{-1}}$ é a matriz diagonal $(a \times a)$ tal que:

$$\mathbf{Diag_r^{-1}} = \begin{bmatrix} \frac{1}{r_1} & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{r_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{r_{a-1}} & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{r_a} \end{bmatrix}$$

Analogamente, designa-se por perfil da coluna j o conjunto das frequências observadas para cada elemento dessa coluna, relativamente ao total de observações nessa coluna. Então, o perfil da coluna j é dado pelos a valores:

$$pc_i^{(j)} = \frac{n_{ij}}{n_{+j}}, \qquad i = 1, \cdots, a$$

Matricialmente, define-se a matriz dos perfis de coluna $\mathbf{P_C}$ por:

$$\mathbf{P_C} = \mathbf{T}\mathbf{Diag}_\mathbf{c}^{-1},$$

onde $\mathbf{Diag}_\mathbf{c}^{-1}$ é a matriz diagonal ($b \times b$) onde os elementos da diagonal são os valores recíprocos do vector de massas de colunas.

Os perfis de linha convergem exactamente num simplex[2]. A nuvem de perfis de linha $\mathbf{N}(A)$ é constituída por a pontos em $\mathbb{R}^b$. Analogamente, a nuvem de perfis de coluna $\mathbf{N}(B)$ é constituída por b pontos em $\mathbb{R}^a$.

No estudo das associações entre as categorias da variável das linhas é importante estudar a dispersão dos pontos da nuvem $\mathbf{N}(A)$. Essa dispersão é medida relativamente ao *centróide* dos perfis de linha, também chamado de centro de gravidade da nuvem dos perfis de linha. O centróide é então o ponto central da nuvem, e relativamente à nuvem $\mathbf{N}(A)$, o seu centróide é dados pelo vector $\mathbf{c}$.

Em relação aos centróides:

- a média ponderada das coordenadas dos a perfis de linha é dada pelo vector:

$$\mathbf{P}_\mathbf{L}^T\mathbf{r} = \mathbf{T}^T\mathbf{Diag}_\mathbf{r}^{-1}\mathbf{r} = \mathbf{c}$$

- a média ponderada das coordenadas dos b perfis de coluna é dada pelo vector:

$$\mathbf{P_C}\mathbf{c} = \mathbf{T}\mathbf{Diag}_\mathbf{c}^{-1}\mathbf{c} = \mathbf{r}$$

O centróide linha de uma tabela de contingência indica geometricamente a posição média dos perfis linha, como se fosse o ponto de equilíbrio da matriz de dados. Na AC o estudo das nuvens é feita com base no seu centróide. Definem-se as distâncias entre perfis linha pela métrica Euclidiana ponderada, chamada métrica ou *Distância Qui-quadrado* (χ^2). A estatística χ^2 indica se há ou não independência entre as linhas e colunas.

$$\begin{aligned} \chi^2 &= \|\mathbf{Diag}_\mathbf{r}^{-1/2}(\mathbf{T} - \mathbf{r}\mathbf{c}^T)\mathbf{Diag}_\mathbf{c}^{-1/2}\|^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n} \end{aligned} \tag{2.18}$$

Sob a hipótese de homogeneidade, a estatística χ^2 pode ser interpretada geometricamente como a diferença dos perfis linha (ou coluna) aos seus respectivos centróides ($\mathbf{c}$ ou $\mathbf{r}$). A

[2]Quando existe dependência linear entre as coordenadas dos vectores perfil (a soma é igual a 1) significa, geometricamente, que os p pontos estão contidos num espaço regular bi-dimensional chamado simplex (Greenacre and Hastie [30]).

significância de χ^2 indica desvios significativos dos perfis linha em relação a seu centróide ou à hipótese de homogeneidade.

Cada perfil de linha está ponderado por um peso proporcional ao respectivo total da linha nos dados originais, correspondentes aos r_i, a massa da linha.

A distância qui-quadrado é usada porque satisfaz o *Princípio da Equivalência Distribucional*. Esta propriedade permite-nos agrupar as linhas ou colunas de $\mathbf{N}$ que são linearmente dependentes, ou seja, estabelece que se dois perfis de linha são equivalentes distributivamente, então essas duas linhas da tabela de contingência podem ser somadas para gerar uma única linha. A nova linha gerada terá o mesmo perfil e a massa igual à soma das massas das linhas somadas, sem que a distância entre os perfis de coluna seja alterada. Do ponto de vista geométrico isso significa que se podem unir dois pontos de uma nuvem que ocupam a mesma posição no simplex, sem provocar alterações noutra nuvem.

A quantidade $\frac{\chi^2}{n}$ é denominada *inércia total* da matriz de dados. A inércia total é a dispersão total da nuvem multidimensional dos perfis (Graffelman [28]).

É possível determinar a inércia ponderada da nuvem de perfis de linha, $\mathbf{N}(A)$ (Cadima [15]):

$$\sum_{i=1}^{a}\sum_{j=1}^{b} r_i \frac{(pl_j^{(i)} - c_j)^2}{c_j} \tag{2.19}$$

Analogamente, a inércia ponderada da nuvem de perfis de coluna, $\mathbf{N}(B)$:

$$\sum_{i=1}^{a}\sum_{j=1}^{b} c_j \frac{(pc_i^{(j)} - r_i)^2}{r_i} \tag{2.20}$$

De Cadima [15] sabe-se então que a inércia ponderada de $\mathbf{N}(A)$ é igual à inércia ponderada de $\mathbf{N}(B)$.

Analogamente à ACP, temos de encontrar o menor subespaço que melhor ajuste a nuvem de pontos. Usando a metodologia aplicada na obtenção dos biplots será possível encontrá-lo.

O problema dual

É possível replicar o mesmo estudo considerando a tabela de contingência transposta $\mathbf{N}_{b\times a}^T$. Os eixos principais dos perfis coluna, ponderada pelas massas, são os elementos de $\mathbf{c}$, num espaço com uma métrica χ^2 definida pela matriz diagonal $\mathbf{Diag}_{\mathbf{r}}^{-1}$. Assim, os elementos de $\mathbf{r}$ e $\mathbf{c}$ desempenham um papel dual, os perfis ponderados por um lado e re-escalando as dimensões por outro lado.

Não é preciso recalcular a solução dual, uma vez que pode ser obtida a partir do primeiro problema. A inércia total e a sua decomposição em inércias principais são exactamente as mesmas nos dois problemas (Greenacre and Hastie [30]). Em cada problema as projecções nos

perfis no seus k-ésimos eixos principais principais podem ser obtidas a partir das projecções dos seus respectivos vértices no problema dual, redimensionando pelo factor igual a $\lambda_k^{1/2}$, a raiz quadrada da k-ésima inércia principal comum (Greenacre and Hastie [30]).

Relações com a Decomposição em Valores Singulares e Biplots

As coordenadas linha e coluna com respeito aos eixos principais podem ser obtidas a partir da SVD da matriz duplamente centralizada e estandardizada:

$$\mathbf{Diag_r}^{-1/2}\left[\mathbf{T}-\mathbf{rc}^T\right]\mathbf{Diag_c}^{-1/2}=\mathbf{U\Sigma V}^T \tag{2.21}$$

Existem diversos critérios de escolha para os marcadores de linhas e marcadores de coluna. Na Secção 2.3.2 está descrita a metodologia AC biplot que iremos usar para os estudos práticos.

2.2.3 Escalonamento Multidimensional

O Escalonamento Multidimensional (*Multidimensional Scaling* - MDS) trata de encontrar a estrutura de um conjunto de medidas de distâncias entre indivíduos e variáveis. Isto é conseguido através da atribuição de observações para posições específicas de um espaço conceitual (geralmente, de duas ou três dimensões), de modo que as distâncias entre os pontos no espaço sejam compatíveis com as dissemelhanças dadas. Em muitos casos, as dimensões no espaço conceitual podem ser interpretadas e utilizadas para melhor compreensão dos dados. Se as variáveis são medidas objectivamente, pode-se usar o MDS como uma técnica para redução dos dados.

Se os dados são dissemelhantes, todas as dissemelhanças devem ser quantitativas e devem-se medir na mesma métrica. Se os dados são multivariados, as variáveis podem ser quantitativas, binárias ou de contagem. A escala de medidas das diferentes variáveis é uma questão importante, pois as diferenças de escala podem afectar a solução. Se as variáveis possuem grandes diferenças na escala de medida, devemos ter em conta a sua caracterização. O MDS é relativamente livre de suposições de distribuições, apenas requer que todas as variáveis estejam, na mesma escala de medida (Álvarez González [35]).

Resumindo, o MDS é usado para descrever qualquer procedimento que parte das "distâncias" entre o conjunto de pontos (indivíduos), ou de qualquer informação sobre essas "distâncias", com o objectivo de chegar a uma configuração de pontos (coordenadas) num espaço de dimensão reduzida (geralmente, 2 ou 3) de modo que a distância euclidiana entre eles reproduza, aproximadamente, as "distâncias originais" no conjunto de dados originais (Álvarez González [35]).

Para reproduzir as dissemelhanças entre n pontos são necessárias no máximo *n-1* dimensões, mas o nosso objectivo é encontrar uma configuração num espaço de dimensão muito

menor, de modo que as distâncias reproduzem as dissemelhanças. Este tipo de representação das dissemelhanças tem uma grande vantagem em qualquer análise, já que poderá mostrar facilmente padrões de comportamento entre os indivíduos, e, nalguns casos, agrupamento de indivíduos (Álvarez González [35]).

Embora tenham surgido na literatura especializada diversas variantes da técnica MDS, os procedimentos de escalonamentos mais usados são as técnicas conhecidas por Escalonamento Clássico e Escalonamento Ordinal. O primeiro é essencialmente um procedimento algébrico para reconstrução das coordenadas dos pontos, assumindo que as diferenças originais são as distâncias euclidianas. O método é robusto mesmo quando as distâncias são distorcidas por erros. Foi originalmente proposto por Torgerson (1952 e 1958), e é também chamado de Escalonamento Métrico, sendo o seu nome mais popular Análise de Coordenadas Principais devido ao trabalho de Gower(1966). Quando os valores das dissemelhanças são subjectivos, sendo mais importante as ordens do que propriamente os valores dessas dissemelhanças, utiliza-se o Escalonamento Ordinal, também conhecido por Escalonamento não métrico, introduzido por Shepard (1962) e Kruskal (1964) (Álvarez González [35]).

O procedimento da Análise em Coordenadas Principais assenta no conceito de matrizes de dissemelhanças e matrizes euclidianas.

Definição 2.2.1. *Uma matriz $\boldsymbol{\Delta}_{n\times n}$ de elemento genérico δ_{ij} é chamada matriz das distâncias ou de dissemelhanças se é simétrica e*

$$\delta_{ii} = 0, \quad \delta_{ij} \geq 0, \quad \textit{quando} \quad i \neq j.$$

Definição 2.2.2. *Seja $\boldsymbol{\Delta}$ uma matriz $n\times n$ de dissemelhanças entre n indivíduos. A matriz $\boldsymbol{\Delta}$, de elemento genérico δ_{ij}, diz-se uma matriz euclidiana se existirem n pontos $\left\{\boldsymbol{x}_{(i)}\right\}_{i=1}^{n} \in \mathbb{R}^p$ tais que*

$$\delta_{ij}^2 = dist^2(\boldsymbol{x}_{(i)} - \boldsymbol{x}_{(j)}) = \|\boldsymbol{x}_{(i)} - \boldsymbol{x}_{(j)}\|^2 = \left(\boldsymbol{x}_{(i)} - \boldsymbol{x}_{(j)}\right)^T \left(\boldsymbol{x}_{(i)} - \boldsymbol{x}_{(j)}\right). \tag{2.22}$$

Da Definição 2.2.2 decorre que uma matriz é euclidiana se for possível determinar n pontos em $\mathbb{R}^p$ tais que a distância euclidiana usual entre qualquer par desses pontos seja a dissemelhança correspondente na matriz $\boldsymbol{\Delta}$. Denotando por $\mathbf{x}_{(i)}$ o vector-linha de $\mathbf{X}$ correspondente a i-ésima linha, pode-se escrever

$$\mathbf{x}_{(i)}^T = \mathbf{e}_i^T \mathbf{X},$$

onde $\mathbf{e}_i$ é o i-ésimo vector da base canónica de $\mathbb{R}^n$. Deste modo, a diferença entre os pontos que representam os indivíduos i e j é dada por:

$$\left(\mathbf{x}_{(i)} - \mathbf{x}_{(j)}\right)^T = \left(\mathbf{e}_i - \mathbf{e}_j\right)^T \mathbf{X} \tag{2.23}$$

e a dissemelhança ao quadrado é:

$$\delta_{ij}^2 = \left(\mathbf{x}_{(i)} - \mathbf{x}_{(j)}\right)^T \left(\mathbf{x}_{(i)} - \mathbf{x}_{(j)}\right) = (\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{X}\mathbf{X}^T (\mathbf{e}_i - \mathbf{e}_j)\,. \tag{2.24}$$

É importante referir que, se uma matriz é euclidiana, os n vectores usados para representar as dissemelhanças em $\mathbb{R}^p$ não são únicos, pelo que a matriz $\mathbf{X}$ também não é única. Por exemplo, se deslocarmos o centro de gravidade dos pontos para a origem das coordenadas, essa transformação deixa distâncias e os ângulos entre os pontos invariantes. Na realidade, qualquer elemento δ_{ij} de uma matriz euclidiana dada por (2.24) também pode ser escrito da seguinte forma:

$$\delta_{ij}^2 = (\mathbf{e}_i - \mathbf{e}_j)^T (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n}) \mathbf{X}\mathbf{X}^T (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n}) (\mathbf{e}_i - \mathbf{e}_j)\,, \tag{2.25}$$

sendo $(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}$ a matriz (de colunas centradas) cujas linhas contêm as coordenadas de cada um dos n pontos na configuração (em torno da origem) procurada. Note que $\mathbf{I}_n$ é a matriz identidade de dimensão $(n \times n)$ e $\mathbf{P}_{\mathbf{1}_n} = \dfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ é a matriz de projecção ortogonal.

Análise de Coordenadas Principais

Dada a matriz das distâncias $\mathbf{\Delta}$, o objectivo do MDS é encontrar os pontos $\ldots P_1, \ldots, P_n \in \mathbb{R}^n$ em $k \leq p$ dimensões tal que, se $\hat{\delta}_{ij}$ denota a distância euclidiana entre P_i e P_j, então a matriz euclidiana $\hat{\mathbf{\Delta}}$ é aproximadamente igual a $\mathbf{\Delta}$. Os pontos P_i são desconhecidos e o valor k também o é. Na prática, $k = 1, 2$ ou 3, a fim de facilitar a interpretação da solução (Mardia et al. [36]). Se $\mathbf{\Delta}$ é uma matriz euclidiana, é possível determinar uma representação euclidiana exacta dos n pontos em p dimensões (ou seja, $k = p$).

Se a matriz de dissemelhanças $\mathbf{\Delta}$ é euclidiana, é possível provar que existe uma relação entre essa matriz $\mathbf{\Delta}$ e a matriz $\mathbf{Q}$ dos produtos internos entre os vectores (centrados) $\mathbf{x}_{(i)}$ definida da seguinte forma:

$$\mathbf{Q} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n}) \mathbf{X}\mathbf{X}^T (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\,. \tag{2.26}$$

Da simetria de $\mathbf{Q}$ e da equação (2.25) resulta que:

$$\delta_{ij}^2 \;=\; \mathbf{e}_i^T\mathbf{Q}\mathbf{e}_i - \mathbf{e}_i^T\mathbf{Q}\mathbf{e}_j - \mathbf{e}_j^T\mathbf{Q}\mathbf{e}_i + \mathbf{e}_j^T\mathbf{Q}\mathbf{e}_j = q_{ii} - 2q_{ij} + q_{jj} \tag{2.27}$$

onde q_{ij} representa o elemento da linha i e coluna j de $\mathbf{Q}$.

Assim, conhecendo a matriz das dissemelhanças $\mathbf{\Delta}$, consegue-se obter $\mathbf{Q}$ a matriz dos produtos internos entre os indivíduos (ver as deduções página 147 de Cadima [15]). Através desta matriz $\mathbf{Q}$, pode-se gerar uma matriz $n \times p$ cujas linhas correspondem ao objectivo da metodologia MDS.

Se a matriz de dissemelhanças $\mathbf{\Delta}$ não é euclidiana é possível obter uma matriz $\mathbf{Q}$ relacionada com $\mathbf{\Delta}$ tal que $\mathbf{Q}$ continua a representar o produto interno entre vectores. Concretamente, se δ_{ij} é o elemento genérico de uma matriz de dissemelhanças $\mathbf{\Delta}$ qualquer, cria-se uma nova matriz $\mathbf{A}$, de elemento genérico definido da forma seguinte:

$$a_{ij} = -\frac{1}{2}\delta_{ij}^2.$$

Em termos matriciais, $\mathbf{A} = -\frac{1}{2}(\mathbf{\Delta} \circ \mathbf{\Delta})$, onde o símbolo "$\circ$" representa o produto de Hadamard[3].

A matriz $\mathbf{Q}$ é definida em termos matriciais, por:

$$\mathbf{Q} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\,\mathbf{A}\,(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\,. \tag{2.28}$$

Para obter os marcadores para os indivíduos no gráfico biplot correspondente ao MDS, aplica-se a SVD à matriz $\mathbf{Q}$.

É somente possível projectar directamente a nuvem de pontos obtida sobre $\mathbb{R}^q$ (geralmente, $q = 2, 3$) retendo apenas as q primeiras colunas da matriz $\mathbf{Y}$, onde:

$$\mathbf{Y}_q = \mathbf{U}_q \mathbf{\Sigma}_q^{1/2},$$

se a matriz $\mathbf{Q}$ for semi-definida positiva.

Se $\mathbf{Q}$ não for semi-definida positiva, alguns dos seus valores próprios serão negativos, o que equivale a dizer que não existe uma representação exacta num espaço euclidiano real. Portanto, não é possível garantir a representação dos indivíduos num espaço $\mathbb{R}^n$ de forma a respeitar as igualdades entre dissemelhanças iniciais e distâncias euclidianas no espaço $\mathbb{R}^n$.

Em suma, uma matriz de dissemelhanças $\mathbf{\Delta}$ é euclidiana no espaço $\mathbb{R}^p$ se e só se a matriz $\mathbf{Q} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})\,\mathbf{A}\,(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}_n})$, com $\mathbf{A} = -\frac{1}{2}(\mathbf{\Delta} \circ \mathbf{\Delta})$, é semi-definida positiva de característica menor ou igual a p.

Quando na matriz $\mathbf{Q}$ existe pelo menos um valor próprio negativo, a matriz de dissemelhanças não é euclidiana. Para contornar este problema, na Análise em Coordenadas Principais existem duas alternativas (Cadima [15]):

1. Se os valores próprios negativos de $\mathbf{Q}$ forem pequenos relativamente à soma dos valores próprios positivos, é possível ignorá-los e trabalhar com uma configuração euclidiana

[3]Designa-se produto de Hadamard de duas matrizes do mesmo tipo, $\mathbf{A}_{n\times p}$, $\mathbf{B}_{n\times p}$, à matriz $\mathbf{C}_{n\times p}$ cujo elemento genérico é dado pelo produto dos correspondentes elementos de $\mathbf{A}$ e $\mathbf{B}$:

$$\mathbf{C}_{n\times p} = \mathbf{A}_{n\times p} \circ \mathbf{B}_{n\times p} \Leftrightarrow c_{ij} = a_{ij}.b_{ij} \quad \forall i \in 1,\ldots,n, \quad \forall j \in 1,\ldots,p.$$

.

aproximada, resultante de considerar apenas os vectores próprios positivos associados de **Q**. Neste caso, escolhe-se o número máximo q de eixos coordenados principais de acordo com as seguintes regras:

Critério do traço: Reter eixos cuja soma de valores próprios associados seja aproximadamente igual ao traço da matriz **Q**.

Critério do valor absoluto: Reter os eixos cujos valores próprios associados sejam maiores do que o módulo do menor valor próprio negativo.

Na prática, é mais frequente escolher $q = 2$ ou 3. Os valores próprios de cada eixo podem ser interpretados como sendo a proporção de variabilidade total explicada pelo eixo. De acordo com Mardia et al. [36], a qualidade da representação obtida utilizando k eixos coordenados principais pode ser medida tomando

(a) $P_1 = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} |\lambda_i|}$;

(b) $P_1 = \frac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{n} \lambda_i^2}$.

2. Outra opção é considerar o problema da constante aditiva. Soma-se uma constante c* a todos os elementos não-diagonais da matriz de dissemelhanças de forma a tornar a matriz **Q** uma matriz semi-definida positiva. Portanto, c* toma o maior valor próprio da seguinte matriz:

$$\begin{bmatrix} \mathbf{0}_n & 2\mathbf{Q} \\ -\mathbf{I}_n & -4\mathbf{Q}(\delta_{ij}) \end{bmatrix}$$

onde $\mathbf{Q}(\delta_{ij})$ corresponde à matriz obtida pela dupla centragem duma matriz análoga à matriz **A**, mas com os elementos dados por $-\frac{1}{2}\delta_{ij}$, em vez de $-\frac{1}{2}\delta_{ij}^2$ (a demonstração pode ser vista em Cailliez [16]).

Para além da Análise de Coordenadas Principais existem outras técnicas de MDS que visam a representação euclidiana dos indivíduos num espaço de dimensão reduzida mas também optimizam um dado critério de qualidade de ajustamento (Cadima [15]). Estas técnicas de Escalonamento Multidimensional podem ser resumidas nos seguintes passos:

1. Fixar a dimensão q do espaço euclidiano onde se deseja a representação.

2. Determinar uma matriz $n \times q$ cujas linhas são as coordenadas de cada indivíduo. Por exemplo, partir da solução q-dimensional produzida pela Análise de Coordenadas Principais.

3. Calcular as distâncias euclidianas usuais entre os n pontos da configuração proposta.

4. Calcular o valor de algum critério de ajustamento que se deseja optimizar.

5. Efectuar alterações à configuração e calcular o novo valor do critério.

6. Repetir o passo anterior até alguma condição de paragem.

O critério de qualidade do ajustamento mais frequentemente usado foi proposto por Kruskal e Shepard, é designado por STRESS e é dado por:

$$STRESS = \sqrt{\frac{\sum_{i=1}\sum_{j<i}(e_{ij} - f(\delta_{ij}))^2}{\sum_i \sum_{j<i} e_{ij}^2}} \tag{2.29}$$

onde δ_{ij} representa a dissemelhança inicial entre os indivíduos i e j, e_{ij} representa a distância euclidiana habitual entre os representantes desses mesmos indivíduos na configuração proposta e f representa uma função crescente[4].

2.2.4 Vantagens e desvantagens

Os três métodos ACP, AC e MDS da Estatística Multivariada têm o mesmo objectivo de redução da dimensionalidade dos dados.

No MDS o objectivo é representar os n indivíduos num espaço euclidiano de modo que as distâncias euclidianas entre cada par dos n pontos sejam iguais, ou o mais próximo possível, às dissemelhanças da matriz original. Se a matriz de dissemelhanças for calculada através das distâncias euclidianas então o problema resume-se a obter as q primeiras CPs. Graficamente, o diagrama de Shepard[5] permite avaliar se as dissemelhanças originais correspondem, ou são aproximadamente iguais, às distâncias euclidianas. Comparativamente à ACP e AC, o MDS é vantajoso no sentido em que é aplicável quando:

- apenas conhecemos a matriz das distâncias euclidianas (ou seja, não é preciso conhecer a matriz de dados iniciais);
- a ACP não é viável, isto é, no caso das dissemelhanças não puderem ser representadas pela matriz das distâncias euclidianas;
- os indivíduos são representados por matrizes, funções ou pelas dissemelhanças entre eles, (como, por exemplo, no caso em que os dados são uma matriz de distâncias rodoviárias).

A maior utilidade da AC é de não precisar validar um modelo ou distribuição dos dados (Benzécri [13]). A AC tem interpretações similares às da ACP mas o tipo de variável é

[4] De acordo com Cadima [15], a escolha duma função decrescente permitirá começar por uma medida de semelhança entre os indivíduos.

[5] Gráfico de dispersão das distâncias euclidianas versus as dissemelhanças dos pontos originais.

diferente, pois na AC as variáveis são categóricas/qualitativas enquanto que na ACP são quantitativas. Além disso, a AC agrupa categorias de linhas e colunas e a ACP agrupa variáveis (da Cunha Jr. [17]). Recentemente, foi publicado um artigo (Greenacre [29]) que permite aplicar a AC a variáveis que não são categóricas nem quantitativas.

2.3 Biplot com redução da dimensionalidade

Um biplot contém, em geral, para além da nuvem de pontos correspondentes aos indivíduos, um conjunto de vectores representativos das variáveis com origem no centróide da nuvem e cujos ângulos verificam:

$$\cos \alpha_{ij} = \cos(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}}$$

Se os vectores formam um cone muito estreito, as variáveis originais são muito correlacionadas. Se eles formam um cone aberto é porque elas são pouco correlacionadas. Falamos então de paralelismo entre as duas variáveis:

$$\text{paralelas} \Leftrightarrow \cos \alpha_{ij} = 1 \Leftrightarrow X_i \text{ e } X_j \text{ são fortemente correlacionadas}$$

e a perpendicularidade entre as variáveis:

$$\text{perpendiculares} \Leftrightarrow \cos \alpha_{ij} = 0 \Leftrightarrow X_i \text{ e } X_j \text{ são não correlacionadas.}$$

Se a nuvem de pontos se estende longitudinalmente ao longo de uma variável experimental e quase não tem projeçcão nas outras, pode interpretar-se que as observações devem a sua variabilidade a essa variável.

2.3.1 ACP Biplot

Vamos apresentar as metodologias usuais para obter um biplot através da ACP.

Consideremos $\mathbf{X}$ a matriz dos dados, $\overline{x}_j$ = média da variável X_j e s_j = desvio padrão da variável X_j.

- SVD sobre a matriz de $\mathbf{D}$ de elemento genérico d_{ij} onde:

$$\begin{array}{ll} d_{ij} = (x_{ij} - \overline{x_j}) & \text{(ACP simples);} \\ d_{ij} = (x_{ij} - \overline{x_j})/\sqrt{n-1} & \text{(ACP simples e corrigida);} \\ d_{ij} = (x_{ij} - \overline{x_j})/s_j & \text{(ACP estandardizada);} \\ d_{ij} = (x_{ij} - \overline{x_j})/(s_j\sqrt{n-1}) & \text{(ACP estandardizada e corrigida).} \end{array}$$

- Coordenadas dos indivíduos: $\mathbf{G} = \mathbf{U\Sigma}$.

- Coordenadas das variáveis: $\mathbf{H} = \mathbf{V}$.
- Representar o biplot.

É possível obter as coordenadas de qualquer variável original em função do espaço das CPs, ou seja, basta calcular $\mathbf{X}^T = \mathbf{L}^{-1}\mathbf{Y}^T$.

A interpretação de um gráfico ACP biplot bidimensional permite detectar relações existentes entre os eixos principais, Y_1 e Y_2, com as variáveis originais $\mathbf{X}$. Se uma variável X_i tem coordenadas baixas num eixo principal (Y_1 ou Y_2) é porque a sua intervenção nesse eixo é relativamente baixa. Contudo, quanto mais altas forem as coordenadas, maior será a representatividade existente. A interpretação das CPs é subjectiva e tem de ser feita de acordo com dados em estudo.

É possível também estabelecer o grau de dependências entre um eixo e uma variável experimental, através dos pesos normalizados que ligam a matriz $\mathbf{X}$ à matriz $\mathbf{Y}$ já que $\mathbf{X}^T = \mathbf{L}^{-1}\mathbf{Y}^T$. Na prática, o grau de dependência é calculado através do coeficiente de correlação linear entre ambas as variáveis:

$$Cor\,(Y_i, X_j) = l_{ij}\frac{\sqrt{Var(Y_i)}}{\sqrt{Var(X_j)}}.$$

Como se pode observar, o coeficiente de correlação depende dos pesos l_{ij}, pelo que podemos concluir que, quanto maior for esse valor, maior será a dependência entre a componente principal e a variável experimental.

Exemplo de aplicação no R

Para exemplificar a ACP biplot, vamos utilizar a base de dados *Iris*, retirada do repositório *UCI* (Fisher [22]), a qual também se encontra incorporada no R. Representaremos os biplots para ACP simples[6].

1. Obter a matriz $\mathbf{D}$.

```
> dados <- read.xlsx("plantas.xlsx", 1)
> caracteristicas<- data.frame(dados)
> iris <- dados[,1:4]
```

2. SVD de $\mathbf{D}$.

   O comando `prcomp` permite efectuar uma ACP através da SVD de uma matriz centrada de dados.

[6] O procedimento completo pode ser consultado no Apêndice A.1.1.

```
iris.pc<-prcomp(iris,center=TRUE,scale=FALSE)
```

- Obtenção das CPs.

```
> iris.pc
Standard deviations:
[1] 2.0554417 0.4921825 0.2802212 0.1538929

Rotation:
                    PC1         PC2         PC3        PC4
comp.sepala  0.36158968 -0.65653988  0.58099728  0.3172545
larg.sepala -0.08226889 -0.72971237 -0.59641809 -0.3240944
comp.petala  0.85657211  0.17576740 -0.07252408 -0.4797190
larg.petala  0.35884393  0.07470647 -0.54906091  0.7511206
```

`Standard deviations` devolve os valores singulares e `Rotation` devolve os vectores singulares à direita. Estes vectores dão-nos os coeficientes da combinação linear das variáveis originais (centradas) que definem cada CP. As CPs são:

$$\begin{aligned} Y_1 &= 0.36158968X_1 - 0.65653988X_2 + 0.58099728X_3 + 0.3172545X_4 \\ Y_2 &= -0.08226889X_1 - 0.72971237X_2 - 0.59641809X_3 + -0.3240944X_4 \\ Y_3 &= 0.85657211X_1 + 0.17576740X_2 - 0.07252408X_3 - 0.4797190X_4 \\ Y_4 &= 0.35884393X_1 + 0.07470647X_2 - 0.54906091X_3 + 0.7511206X_4 \end{aligned}$$

onde X_1 representa o comprimento da sépala, X_2 representa a largura da sépala, X_3 representa o comprimento da pétala e X_4 representa a largura da pétala. Os valores obtidos, chamados *scores* de cada observação na CP, são apenas apresentados quando solicitados:

```
> iris.pc$x
                PC1         PC2          PC3           PC4
  [1,] -2.68420713 -0.32660731  0.021511837  1.006157e-03
  [2,] -2.71539062  0.16955685  0.203521425  9.960242e-02
  [3,] -2.88981954  0.13734561 -0.024709241  1.930454e-02
                   ...
[149,]  1.90162908 -0.11587675 -0.722873561  4.087282e-02
[150,]  1.38966613  0.28288671 -0.362317832 -1.563104e-01
```

- Visualização da proporção de variância de cada uma das CPs e a proporção acumulada destas.

```
> summary(iris.pc)
Importance of components:
                          PC1   PC2    PC3     PC4
Standard deviation     2.055 0.492 0.2802 0.15389
Proportion of Variance 0.925 0.053 0.0172 0.00518
Cumulative Proportion  0.925 0.978 0.9948 1.00000
```

O comando anterior permite visualizar a qualidade de representação de cada CP. Analisando a linha referente à proporção acumulada, verifica-se que a primeira CP explica aproximadamente 92,5% da variabilidade dos dados. As duas primeiras componentes acumularam uma percentagem bastante satisfatória, com aproximadamente 97,8% da variabilidade dos dados. As demais componentes absorvem apenas cerca de 2,238% da variabilidade, portanto, as novas variáveis Y_1 e Y_2 podem substituir as quatro variáveis originais com pouca perda de informação. Através da Figura 2.2(a), obtida usando

```
> screeplot(iris.pc,type="l",main="Screeplot de iris.pc")
```

verificamos que a regra do cotovelo sugere que se podem usar as duas primeiras CPs para representar os dados.

- Visualização gráfica dos dados e existência de correlações:

```
> plot(iris.pc$x[,1:2],
+ col=rep(especies),
+ main="Plot das duas 1ªs Componentes Principais",
+xlab="1ª Componente Principal
+ (92.5%)", ylab="2ª Componente Principal (5.3%)",
+ type="p", pch=19)
> legend(0,-0.8,c("Iris setosa","Iris versicolor","Iris virginica"),
+ col=c(1,2,3), text.col="black",bg='gray90',pch=19)
> abline(h=0,lty=2); abline(v=0,lty=2)
```

Dos comandos acima, obtivemos a Figura 2.2(b) onde se encontra a nuvem dos 150 indivíduos no espaço definido pelas duas primeiras CPs.

Olhando para a matriz das covariâncias, observe-se que a variável X_3 (*comp.petala*) tem uma variabilidade muito superior as restantes variáveis:

```
> var(iris)
             comp.sepala larg.sepala comp.petala larg.petala
comp.sepala   0.68569351 -0.03926846   1.2736823   0.5169038
larg.sepala -0.03926846  0.18800403  -0.3217128  -0.1179812
```

```
comp.petala  1.27368233 -0.32171275   3.1131794   1.2963875
larg.petala  0.51690380 -0.11798121   1.2963875   0.5824143
```

Por isso, essa variável é a que mais contribui para a primeira CP. Podemos confirmar este resultado analisando os coeficientes de correlação da primeira CP com cada uma das variáveis:

```
> cor(iris,prcomp(iris)$x)
                   PC1         PC2         PC3         PC4
comp.sepala  0.8975449 -0.39023141  0.19661200  0.05896054
larg.sepala -0.3899934 -0.82831259 -0.38545012 -0.11502877
comp.petala  0.9978541  0.04903006 -0.01151811 -0.04184113
larg.petala  0.9664842  0.04818017 -0.20160693  0.15146498
```

Confirmamos então que a correlação da variável X_3 (*comp.petala*) com a primeira CP é muito elevada e observe-se também que a correlação da variável X_4 (*larg.petala*) com a primeira CP é muito elevada. Tal facto advém dessas duas variáveis estarem fortemente correlacionadas e podemos confimar olhando para a matriz das correlações entre as variáveis originais:

```
> cor(iris)
             comp.sepala larg.sepala comp.petala larg.petala
comp.sepala    1.0000000  -0.1093692   0.8717542   0.8179536
larg.sepala   -0.1093692   1.0000000  -0.4205161  -0.3565441
comp.petala    0.8717542  -0.4205161   1.0000000   0.9627571
larg.petala    0.8179536  -0.3565441   0.9627571   1.0000000
```

3. Representação do biplot.

- A escolha de α (da equação (2.5)) é identificada dentro do comando `biplot` no parâmetro `scale`[7] (alp [1]). O biplot pode ser obtido automaticamente (Figura 2.2(c)) do modo seguinte:

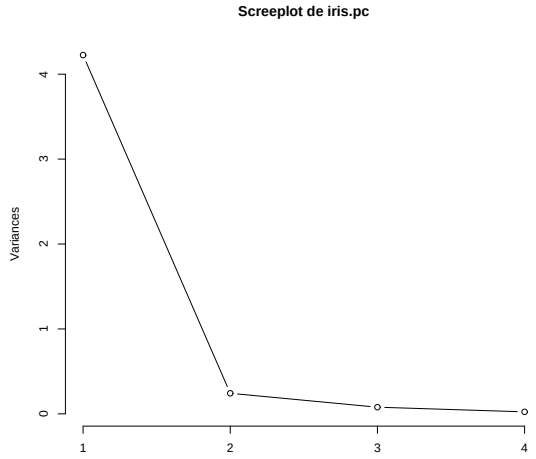
```
> biplot(iris.pc,choices=1:2,main="Biplot das 2 1ªs CP",
+xlab="1ª CP", ylab="2ª CP",
+ var.axes=TRUE,scale=0,pc.biplot=TRUE)
> abline(h=0,lty=2)
> abline(v=0,lty=2)
```

[7]Os marcadores dos indivíduos são dados por $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}^{\alpha}$ com $\alpha = 1$. No R, $\mathbf{\Sigma}^{\alpha}$ é denotado por $lambda^{1-scale}$. Assim, para tomar $\alpha = 1$, no R deverá ser tomado $scale = 0$.

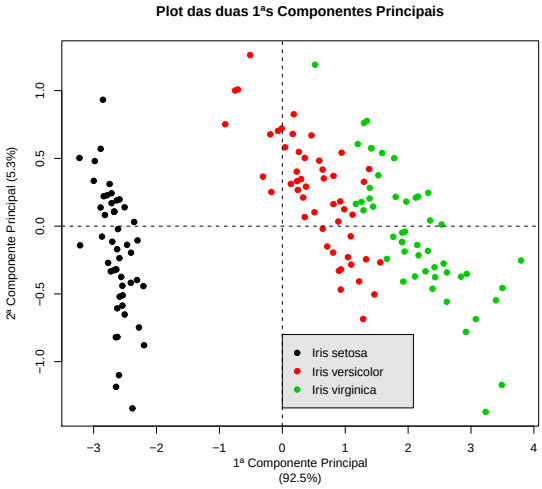
- Obtenção do biplot das duas primeiras CPs sem recurso ao comando `biplot` (Figura 2.2(d)).

```
> N<-nrow(iris)
> setosa<-rep(1,50); versicolor<-rep(2,50); virginica<-rep(3,50)
> grupos<-c(setosa,versicolor,virginica)
> acp<-prcomp(iris,scale=FALSE, center=TRUE)
> xmin <- min(acp$x[,1])
> xmax <- max(acp$x[,1])
> ymin <- min(acp$x[,2])
> ymax <- max(acp$x[,2])
> plot(c(xmin,xmax),c(ymin,ymax),col="white",
+xlab="1ª Componente Principal(92.5%)",
+  ylab="2ª Componente Principal (5.3%)")
> text(iris.pc$x[,1],iris.pc$x[,2],1:N,col=grupos)
> title("PCA Biplot manual das iris para as duas primeiras CP")
> legend(2,1.2,c("Iris setosa","Iris versicolor",
+"Iris virginica"),  col=c(1,2,3),
+text.col="black",bg='gray90',pch=19)
> abline(v=0,lty=2); abline(h=0, lty=2)
> xl.min = min(0,min(acp$rotation[,1]))
> xl.max = max(0,max(acp$rotation[,1]))
> yl.min = min(0,min(acp$rotation[,2]))
> yl.max = max(0,max(acp$rotation[,2]))
> xl.scale = max(abs(xmax),abs(xmin))/max(abs(xl.max),
+abs(xl.min))*0.75
> yl.scale = max(abs(ymax),abs(ymin))/max(abs(yl.max),
+abs(yl.min))*0.75
> arrows(rep(0,100),rep(0,100),acp$rotation[,1]*xl.scale,
+ acp$rotation[,2]*yl.scale, col="blue")
> text(acp$rotation[,1]*xl.scale, acp$rotation[,2]*yl.scale ,
+ colnames(iris))
```

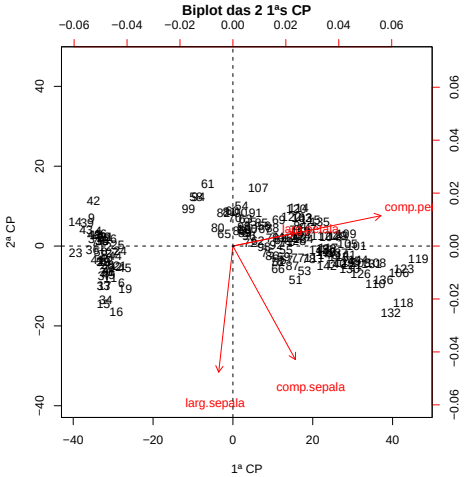
Este conjunto de instruções possibilitou a construção do biplot onde as espécies estão discriminadas por cor, atributo não disponível no comando `biplot` (confrontar Figuras 2.2 (c)-(d)).



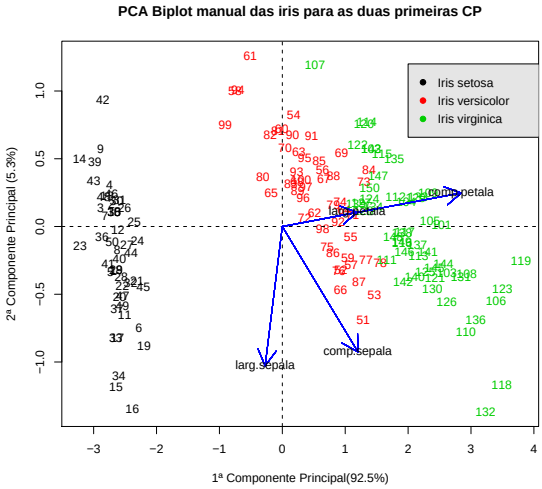
(a) Screeplot das CPs.



(b) Nuvem de pontos no espaço definido pelas duas primeiras CPs.



(c) Biplot das duas primeiras CPs usando o comando `biplot`.



(d) Biplot das duas primeiras CPs sem recurso ao comando `biplot`.

Figura 2.2: Gráficos obtidos através da ACP biplot para os dados *Iris*.

Das Figuras 2.2 (b) e (d) observe-se que as três espécies aparecem em grupos relativamente bem separados. Evidencia-se o elevado grau de correlação entre as duas variáveis associadas às pétalas (as setas associadas aos respectivos marcadores apontam no mesmo sentido). A variância da variável X_3 (*comp.petala*) é maior do que as restantes (visível no maior comprimento do respectivo marcador). Este tipo de conclusão é extensível entre marcadores de variáveis e eixos coordenados (CPs 1 e 2). O facto de as setas, que servem de marcadores das

variáveis correspondentes às pétalas, serem quase horizontais reflecte a elevada correlação dessas variáveis com a primeira CP. Analogamente, o facto de a seta, que serve de marcador da variável correspondente à largura das sépalas, ser quase vertical reflecte a elevada correlação dessa variável com a segunda CP. Observe-se também que todos os indivíduos da espécie *Iris setosa* têm as pétalas mais pequenas porque a projecção ortogonal dos respectivos marcadores de indivíduos na direcção dos marcadores X_4 (*larg.petala*) e X_3 (*comp.petala*) fica abaixo da média. Analogamente, os indivíduos 119, 123 e 106 aparentam ter as maior medições de pétalas.

2.3.2 AC Biplot

Consideremos $\mathbf{X}$ uma matriz de dados brutos de elemento genérico x_{ij} com todos os $x_{ij} > 0$. De acordo com (Pineda-Vargas et al. [45]) e (Greenacre [29]), a obtenção de biplots associados à AC resulta do seguinte procedimento:

- Obter a matriz $\mathbf{T}$: $t_{ij} = \frac{x_{ij}}{x_{tot}}$, onde $x_{tot} = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}$.
- Obter os vectores de massa para as linhas e colunas:

$$\begin{aligned} \mathbf{r} &= \mathbf{T}\mathbf{1_n} \\ \mathbf{c} &= \mathbf{T}^T\mathbf{1_p} \end{aligned}$$

- Calcular:

$$\mathbf{Diag_r}^{-1/2}(\mathbf{T} - \mathbf{r}\mathbf{c}^T)\mathbf{Diag_c}^{-1/2};$$

  onde $\mathbf{Diag_r}$ e $\mathbf{Diag_c}$ são as matrizes diagonais calculadas a partir dos vectores de massas $\mathbf{r}$ e $\mathbf{c}$, respectivamente.

- SVD de $\mathbf{Diag_r}^{-1/2}(\mathbf{T} - \mathbf{r}\mathbf{c}^T)\mathbf{Diag_c}^{-1/2}$.
- Coordenadas principais para os indivíduos: $\mathbf{F} = \mathbf{Diag_r}^{-1/2}\mathbf{U}\mathbf{\Sigma}$.
- Coordenadas padronizadas para as variáveis: $\mathbf{\Phi} = \mathbf{Diag_c}^{-1/2}\mathbf{V}$.
- Representar o biplot.

Em Villalobos Aguayo [51] são apresentados resultados sobre relações das posições dos pontos no plano definido pelas coordenadas principais:

1. Se duas linhas têm uma estrutura semelhante, ou seja, se os seus perfis são similares, então ambas as linhas relacionam-se de modo similar com as colunas. No gráfico, os seus pontos estarão localizados próximos um do outro.

2. Se dois pontos de linha estão muito distanciados, então as suas linhas estão relacionadas de modo diferentes com as colunas.

3. Quando dois pontos de linha estão em direcções opostas à origem, eles desviam-se de forma oposta do perfil médio.

Exemplo de aplicação no R

Vamos utilizar novamente a base de dados *Iris*, para representar o biplot obtido através da AC[8].

Tendo em conta que os dados *Iris* são referentes a comprimentos, podemos logo realizar a AC sem prévia transformação dos dados. Na ACP, a escolha do número de eixos depende da proporção da variância acumulada. No caso da AC, o número de eixos a serem seleccionados depende das percentagens acumuladas das inércias principais. No R, efectuando o seguinte conjunto de comandos:

```
> library(rgl) #para poder utilizar a package ca
> library(ca)
> ca.novo <- ca(iris)
> summary(ca(iris))
```

obtem-se:

```
Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.061115  95.5  95.5  ************************
 2      0.002228   3.5  99.0  *
 3      0.000623   1.0 100.0
        -------- -----
 Total: 0.063966 100.0


Rows:
      name   mass  qlt  inr    k=1  cor ctr    k=2 cor ctr
1   |    1 |    5 1000   15 | -448 1000  16 |   -3   0   0 |
2   |    2 |    5  994   12 | -407  985  12 |   39   9   3 |
3   |    3 |    5 1000   14 | -441 1000  14 |   -3   0   0 |
                        ...
```

[8] O procedimento completo pode ser consultado no Apêndice A.1.2.

```
149 |  149 |    8  997    6 |  182  764   4 | -100 233  38 |
150 |  150 |    8  968    3 |  162  931   3 |  -32  37   4 |

Columns:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 | cmps |  422  960   79 | -102 880  72 |   31  79 179 |
2 | lrgs |  220  994  334 | -306 968 339 |  -50  26 249 |
3 | cmpp |  271  991  317 |  271 980 325 |   28  11  97 |
4 | lrgp |   87  994  270 |  432 933 264 | -111  61 475 |
```

Observamos que uma representação gráfica dos indivíduos em duas dimensões apresenta uma percentagem acumulada das inércias principais de 99%. Portanto, as conclusões tiradas na análise das observações projectadas no espaço reduzido serão muito precisas.

O biplot para AC, utilizando o método das coordenadas principais para as linhas e com as respectivas massas associadas, pode ser obtido directamente do R (Figura 2.3(a)) usando a seguinte instrução:

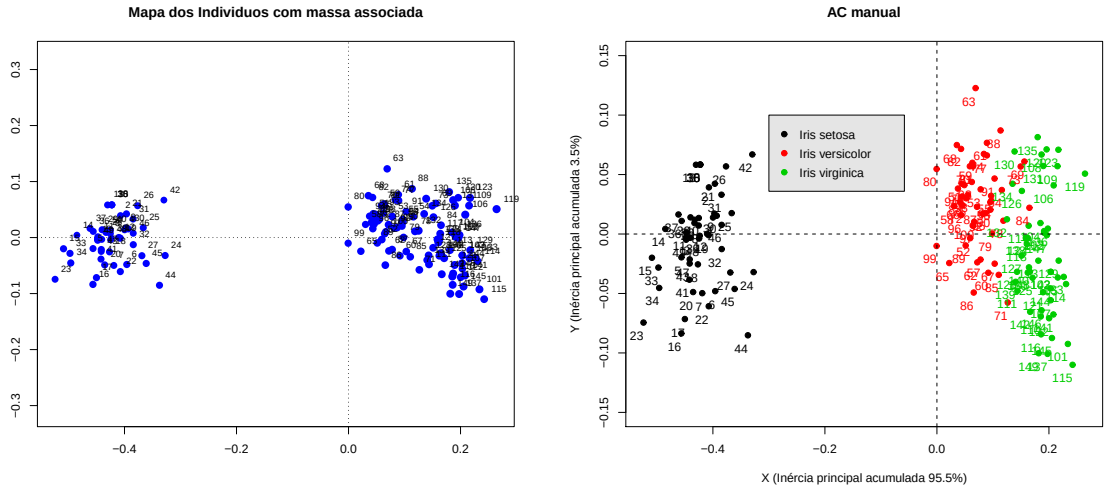
```
> plot(ca(elementos), map="rowprincipal", what=c("all","none"),
+ mass=c(TRUE,FALSE),labels=c(2,0)
+ ,main="Mapa dos Individuos com massa associada")
```

Analogamente ao que fizemos para a ACP biplot, segue um conjunto de instruções para a construção de AC biplots com maior flexibilidade na discriminação das espécies por cor (Figura 2.3(b)):

```
> P<- iris/sum(iris)
> rm<-apply(P,1,sum)
> cm<-apply(P,2,sum)
> Q<-P-rm%*%t(cm)
> Dr<-as.matrix(diag(sqrt(1/rm)))
> Dc<-as.matrix(diag(sqrt(1/cm)))
> S<-Q/sqrt(rm%*%t(cm))
> s<-svd(S)
> U<-s$u
> V<-s$v
> Sigma<-diag(s$d)
> F<-Dr %*% U %*% Sigma
> ca.novo.raw<-ca.novo
> ca.novo.raw$rowcoord<-F
```

```
> x<-ca.novo.raw$rowcoord[,1]
> y<-ca.novo.raw$rowcoord[,2]
> plot(x,y,pch=19,col=rep(especies),xlab="X", ylab="Y",ylim=c(-0.15,0.15))
> title("AC manual")
> text(x,y,lab=labels(row.names(iris)),1:N,col=grupos)
> abline(v=0,lty=2); abline(h=0, lty=2)
> legend(-0.2,.10,c("Iris setosa","Iris versicolor","Iris virginica"),
+ col=c(1,2,3), text.col="black",bg='gray90',pch=19)
```

Olhando para Figura 2.3(a), podemos distinguir dois grupos. Quase todas as observações têm massa igual, ou seja, a importância relativa de cada observação é semelhante às outras todas. Através do biplot da Figura 2.3(b) conseguimos distinguir as três espécies de iris, mas tal já acontecia na ACP. Algumas observações correspondentes à espécie *Iris versicolor* estão mais próximas e, neste caso, agrupadas com as observações correspondentes à espécie *Iris virginica*.



(a) AC biplot com recurso ao comando `plot(ca(···))` para as coordenadas principais dos indivíduos e massas associadas.

(b) AC biplot sem recurso ao comando `plot(ca(···))` para as coordenadas principais dos indivíduos.

Figura 2.3: Gráficos obtidos através da AC biplot para os dados *Iris*.

2.3.3 MDS Biplot

Seja então $\mathbf{\Delta}$ uma matriz de dissemelhanças simétricas, com elementos δ_{ij} não negativos, sobre a qual não sabemos se é euclidiana. Para a obtenção de biplots associados ao MDS aplica-se o seguinte algoritmo:

- Calcular $\mathbf{A}$ tal que
$$\mathbf{A} = -\frac{1}{2}(\mathbf{\Delta} \circ \mathbf{\Delta})$$
- Calcular $\mathbf{Q}$ tal que
$$\mathbf{Q} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1}n})\mathbf{A}(\mathbf{I}_n - \mathbf{P}_{\mathbf{1}n})$$
- SVD de $\mathbf{Q}$
$$\mathbf{Q} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T.$$
- Se $\mathbf{Q}$ for semi-definida positiva:
  - Calcular $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}^{1/2}$
  - Tomar $\mathbf{Y}_q = \mathbf{U}_q\mathbf{\Sigma}_q^{1/2}$.
- Se $\mathbf{Q}$ não for semi-definida positiva, seguir o critério do problema da constante aditiva, o qual consiste em

  Somar c^* a todos os elementos não diagonais de $\mathbf{\Delta}$ de forma a tornar $\mathbf{Q}$ semi-definida positiva. Usar c^* igual ao maior valor próprio da matriz:
$$\begin{bmatrix} \mathbf{0}_n & 2\mathbf{Q} \\ -\mathbf{I}_n & -4\mathbf{Q}(\delta_{ij}) \end{bmatrix}$$
- Representar o biplot.

Exemplo de aplicação no R

Vamos utilizar novamente a base de dados *Iris* para representar o biplot obtido através do MDS[9].

Inserindo no R as instruções:

```
> delta<-as.matrix(dist(iris),150,150)
> delta.mds<- cmdscale(delta, k=149, eig=TRUE)
```

obtemos

```
Warning messages:
1: In cmdscale(delta, k = 149, eig = TRUE) :
  some of the first 149 eigenvalues are < 0
2: In sqrt(ev) : NaNs produced
```

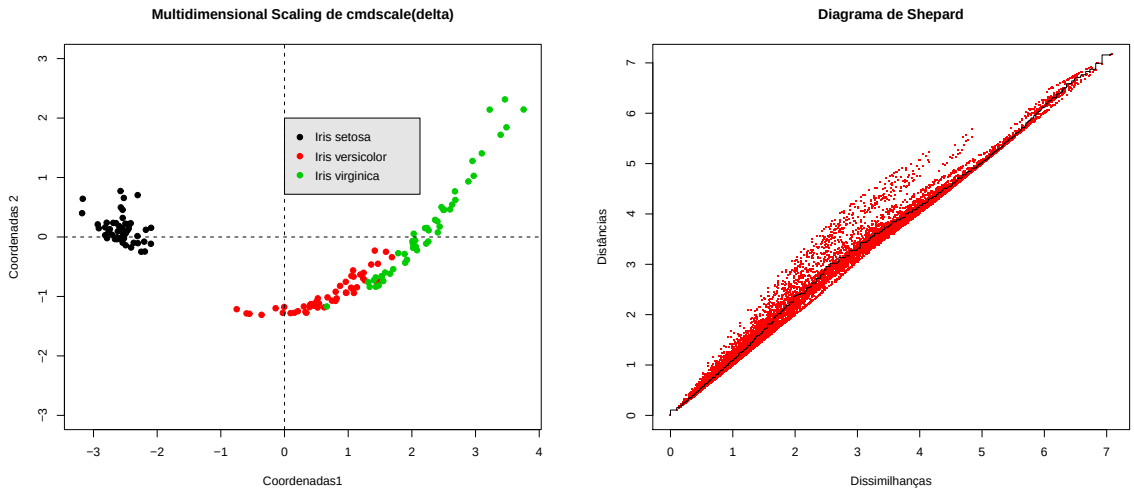
O programa informa que a matriz $\mathbf{Q}$ não é semi-definida positiva, pelo que é necessário resolver esse problema através do Problema da constante aditiva do seguinte modo (Cadima [15] e no R hel [3]):

[9] O procedimento completo no R pode ser consultado no Apêndice A.1.3.

```
> delta.mds<- cmdscale(delta, k=149, eig=TRUE, add=TRUE)
```

Para o gráfico biplot (Figura 2.4(a)) fazemos:

```
> x<-delta.mds$points[,1]
> y<-delta.mds$points[,2]
> plot(x, y, xlab="Coordenadas1",
+ ylab="Coordenadas 2",pch=19,col=rep(especies),ylim=c(-3,3))
> title("Multidimensional Scaling de cmdscale(delta)")
> legend(1,2,c("Iris setosa","Iris versicolor","Iris virginica"),
+ col=c(1,2,3), text.col="black",bg='gray90',pch=19)
> abline(v=0,lty=2); abline(h=0, lty=2)
```



(a) MDS biplot com recurso ao comando `cmdscale(···)`.

(b) Diagrama de Shepard para os dados *Iris*.

Figura 2.4: Gráficos obtidos através do MDS biplot para os dados *Iris*.

A análise da Figura 2.4(a), relativa ao MDS biplot obtido para estes dados da iris, vem reforçar a ideia de que as observações correspondentes à espécie *Iris setosa* estão muito mais distantes e dissemelhantes do que as outras duas espécies. A Figura 2.4(b) é denominada de diagrama de Shepard e permite visualizar as discrepâncias existentes entre as dissemelhanças originais e as dissemelhanças no espaço reduzido definido pela Análise de Coordenadas Principais. Este diagrama de Shepard foi obtido usando as seguintes instruções:

```
> deltaSh<-Shepard(delta[lower.tri(delta)],delta.mds$points)
> plot(deltaSh,pch=".",xlab="Dissimilhanças",
+ ylab="Distâncias", xlim= range(deltaSh$x),
+ ylim=range(deltaSh$x),col="red")
```

```
> title("Diagrama de Shepard")
> lines(deltaSh$x,deltaSh$yf,type="S")
```

Podemos calcular uma medida de qualidade de ajustamento:

```
> sum((delta.mds$eig[1:2]))/sum((delta.mds$eig)^2)
[1] 0.001766257
```

Ao contrário das duas outras análises (ACP e AC) esta representação (MDS) é relativamente fraca tendo em conta o resultado da medida de qualidade acima calculada. Tal é visível no diagrama de Shepard onde se nota um relativo afastamento dos pontos em relação à bissectriz (Figura 2.4(b)).

Capítulo 3

Estudo Experimental

3.1 Matrizes de dados

Neste capítulo, vamos aplicar os métodos explicados no capítulo anterior a dois conjuntos de dados reais (não académicos). Começaremos por descrever com maior detalhe os conjuntos Dados 1 e Dados 2 introduzidos no Capítulo 1.

3.1.1 Dados 1: Dados de *microarrays*

Este conjunto de dados foi retirado do site (Alon [11]) contendo em diversos ficheiros, toda a descrição da experiência de *microarray* elaborada para o estudo da diferenciação do nível de expressão genética entre células cancerosas e normais do colón de 40 indivíduos. Destacamos aqui os ficheiros I2000, *names* e *tissues*.

O ficheiro I2000 é constituído pelos níveis de expressão genética de dois mil genes de interesse relativo aos sessenta e dois tecidos. Os genes estão colocados por ordem decrescente de intensidade mínima não tendo sido processada qualquer tipo de normalização sobre os dados.

O ficheiro *names* contém a identificação (por etiquetas) e a descrição de cada um dos genes, na mesma ordem que estão dispostos no ficheiro I2000.

O ficheiro *tissues* contém a identidade dos sessenta e dois tecidos amostrais dada por um número e um sinal. Os números correspondem à identificação do paciente e o sinal corresponde ao tipo de tecido (sinal positivo: corresponde a um tecido normal e o sinal negativo: a um tecido canceroso).

Com base na informação recolhida nos ficheiros I2000, *names* e *tissues*, construímos numa folha de cálculo do Excel a matriz de dados $\mathbf{X}_{62\times 2000}$ relativa ao conjunto Dados 1 para análise. Esta pode ser representada do seguinte modo:

indivíduo	gene1 gene2 $\cdots$ gene1999 gene2000
T1	
N1	
$\vdots$	nível de expressão genético
N39	
T40	
N40	

Trata-se de uma matriz de dados constituída por 2000 variáveis, em que cada variável representa um gene, e por 62 indivíduos, representando as 62 amostras de tecido analisadas, identificados por uma letra e um número. Os números correspondem à identificação dos pacientes e as letras ao tipo de célula: se for "T" a célula é cancerosa, se for "N" a célula é normal.

3.1.2 Dados 2: Dados de pares de codões nas sequências de DNA

Este conjunto de dados foi construído com base na leitura das sequências completas de DNA de 123 espécies (listadas no Apêndice B).

A partir do Anaconda é possível fazer corresponder a sequência genómica de qualquer espécie numa tabela 61[1] $\times$ 61 de resíduos ajustados de Pearson[2] resultante do teste χ^2 para a independência em pares de codões consecutivos - ver Tabela 1.1. Esquematicamente, uma vez que esses resíduos dependem do tamanho n do genoma da espécie (Pinheiro et al. [46]), para efectuar comparações entre o genoma de diferentes espécies, os valores d_{ij} da Tabela 1.1 foram divididos pela raiz quadrada do tamanho do genoma, obtendo-se uma nova tabela 61×61 de resíduos ajustados normalizados, $\frac{d_{ij}}{\sqrt{n}}$.

No presente trabalho propomos explorar o comportamento do grau de preferência de pares de codões consecutivos das 123 espécies usando biplots. Para tal, a tabela de resíduos normalizados $\frac{d_{ij}}{\sqrt{n}}$ de cada espécie foi transformada num vector 1×3721 ($3721 = 61 \times 61$) e, a seguir, os 123 vectores assim obtidos foram condensados numa única matriz. Deste modo obtivemos a matriz de dados 123 indivíduos $\times$ 3721 variáveis a ser usada para o nosso estudo (Dados 2). A Tabela 3.1 contém a informação completa do conjunto de dados Dados 2 para análise. Observe-se que os indivíduos (espécies) estão repartidos em cinco reinos: *Animalia* com setenta espécies, *Monera* com trinta espécies, *Fungi* com onze espécies, *Plantae* que contém oito espécies e o reino *Protista* com quatro espécies. Para além dos respectivos reinos,

[1] Uma vez que estamos a estudar o grau de preferência de cada par de codão, não faz sentido incluir os codões de finalização. Estes foram retirados da análise (Tabela 1.1)

[2] Os resíduos ajustados de Pearson são denotados por d_{ij} em Pinheiro et al. [46].

nome das espécies e resíduos ajustados normalizados, a cada indivíduo foi-lhe atribuído uma etiqueta abreviada para simplificar a sua identificação nas representações gráficas.

Reino	Abreviatura	Espécie	AAA-AAA $\cdots$ UUU-UUU
$Animalia_1$			
$\vdots$			
$Animalia_{70}$			
$Monera_1$			
$\vdots$			
$Monera_{30}$			
$Fungi_1$			
$\vdots$			$\frac{(d_{ij})\text{espécie}_k}{\sqrt{n_k}}$
$Fungi_{11}$			
$Plantae_1$			
$\vdots$			
$Plantae_8$			
$Protista_1$			
$\vdots$			
$Protista_4$			

Tabela 3.1: Tabela representativa do ficheiro de dados relativos ao conjunto Dados 2.

Por conseguinte, a matriz de dados $\mathbf{X}_{123\times3721}$ é constituída por 3721 variáveis e por 123 indivíduos, onde cada variável representa um possível par de codões e cada indivíduo representa uma espécie.

3.2 Resultados e Análise

Começamos por proceder à leitura de cada ficheiro no R. Para tal, foi necessário, em primeiro lugar, verificar se o programa Java já se encontra instalado no computador uma vez que é um requisito para podermos usar as *packages* de leitura consideradas no nosso estudo. Os procedimentos completos para aceder aos ficheiros Dados 1 e Dados 2 e facultar a sua leitura estão disponíveis no Apêndice A.2 e A.3, respectivamente.

3.2.1 Dados 1: Dados de *microarrays*

Uma vez lido o ficheiro Dados 1, aplicámos as técnicas ACP, AC e MDS (cada uma separadamente) com vista a reduzir o espaço das 2000 variáveis a um plano (dimensão 2) e, consequentemente, projectar as 62 observações naquele espaço de dimensão 2. O objectivo

é investigar relações entre genes e/ou amostras de tecidos usando os biplots e confirmar os resultados obtidos com a aplicação das três técnicas de redução de dados ACP, AC e MDS realizada por Park et al [44].

Começamos por aplicar ACP. Os comandos do R utilizados para a ACP encontram-se na Secção A.2.1. A AC simples foi feita com auxílio da *package* `ca` implementada por Nenadic e Greenacre em 2007. Para compreendermos a estrutura e funcionamento da *package*, consultámos Nenadic and Greenacre [38] e *helps* do R (Nenadic and Greenacre [40] e Nenadic and Greenacre [39]). Todos os comandos utilizados para AC encontram-se na Secção A.2.3. Por fim, aplicámos o MDS, e os comandos do R utilizados encontram-se na Secção A.2.4.

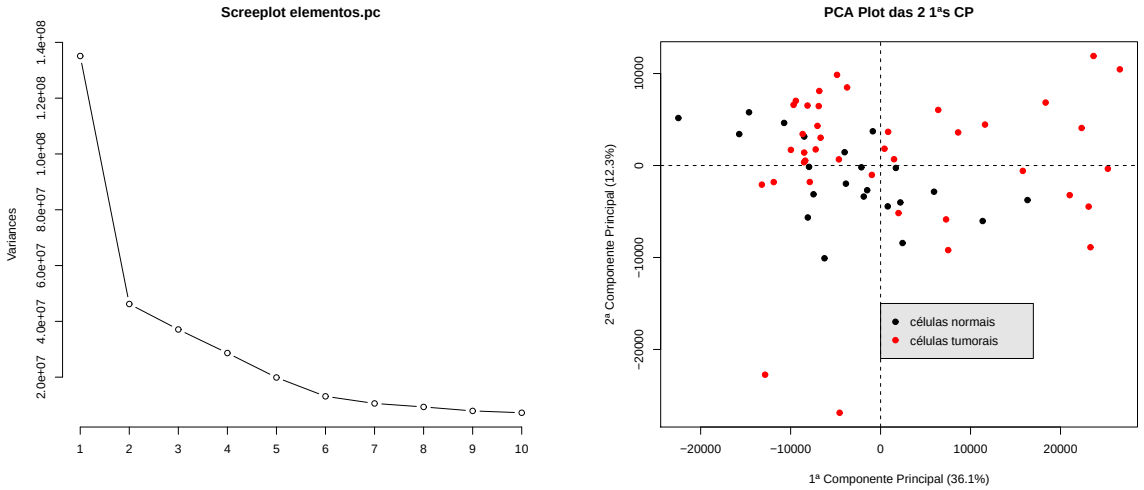
Ao aplicar ACP, começamos por calcular a proporção de variância explicada pelas novas variáveis (CP) obtendo:

```
Importance of components:
                            PC1      PC2      PC3      PC4      PC5      PC6
Standard deviation     1.16e+04 6798.677 6090.100 5.35e+03 4.46e+03 3.62e+03
Proportion of Variance 3.61e-01    0.123    0.099 7.65e-02 5.31e-02 3.51e-02
Cumulative Proportion  3.61e-01    0.484    0.584 6.60e-01 7.13e-01 7.48e-01
                            PC7      PC8      PC9     PC10     PC11     PC12
Standard deviation     3.25e+03 3.05e+03 2.81e+03 2.69e+03 2.30e+03 2.21e+03
Proportion of Variance 2.83e-02 2.49e-02 2.11e-02 1.93e-02 1.42e-02 1.31e-02
Cumulative Proportion  7.76e-01 8.01e-01 8.22e-01 8.42e-01 8.56e-01 8.69e-01
                                    ...
                           PC61     PC62
Standard deviation     272.9597 5.16e-12
Proportion of Variance   0.0002 0.00e+00
Cumulative Proportion    1.0000 1.00e+00
```

Analisando a linha referente à proporção acumulada, verifica-se que a primeira CP explica aproximadamente 36,1% da variabilidade dos dados e as duas primeiras componentes acumulam aproximadamente 48,4% de explicação da variabilidade dos dados. As demais componentes absorvem apenas cerca de 51,6% da variabilidade. Tendo em conta que se trata de projectar as observações de um espaço 2000-dimensional num espaço 2-dimensional, parece aceitável assumir aquela perda de informação e tomar as duas primeiras CPs como as novas direcções do espaço reduzido onde se projectarão as 62 observações.

Focando a atenção no screeplot da Figura 3.1(a), podemos observar dois cotovelos, um em "2" e outro em "6" sugerindo a escolha de 2 ou 6 CPs. Assim, prosseguiremos a análise tomando o subespaço bi-dimensional de $\mathbb{R}^{62}$ gerado pelas duas primeiras CPs em vez de seis,

pois o nosso objectivo é a visualização gráfica das observações num plano usando biplots.



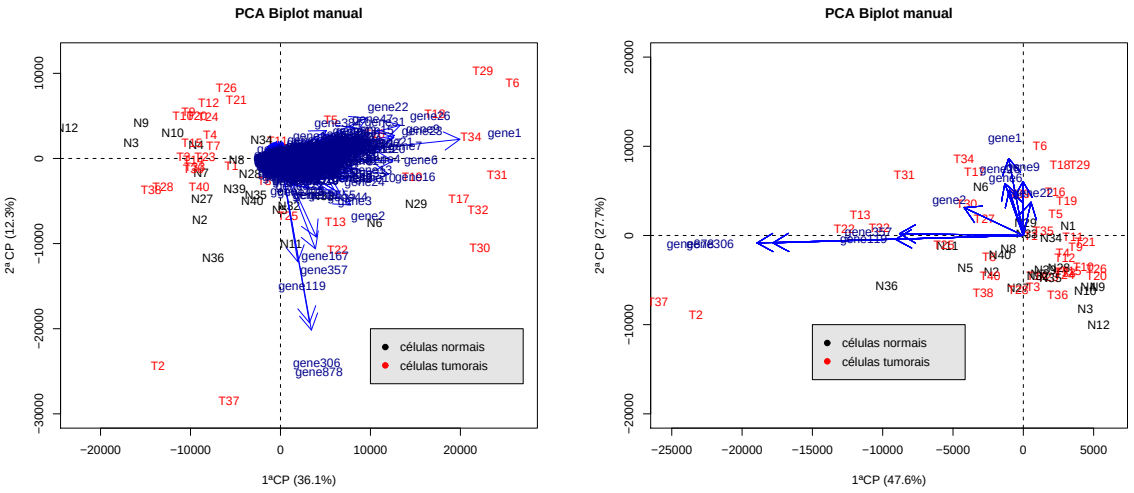
(a) Screeplot das Componentes Principais para Dados 1.

(b) Nuvem de pontos no espaço definido pelas duas primeiras CPs.

Figura 3.1: Gráficos obtidos através da ACP para o conjunto Dados 1.

A imagem reproduzida na Figura 3.1(b) corresponde então a uma representação aproximada da nuvem de pontos originais, com uma percentagem de variabilidade dos dados originais explicada em 48,4%. Na Figura 3.2(a) representamos o biplot (manual) da ACP. Podemos observar uma correlação relativamente forte entre os genes 306, 878 e 119, uma vez que as setas associadas aos respectivos marcadores apontam no mesmo sentido. Podemos também observar que a variância das variáveis gene 878, 306 e 1 é maior do que as restantes (situação visível no maior comprimento do respectivo marcador). Nota-se que este tipo de conclusão é extensível à relação entre marcadores de variáveis e eixos associados às variáveis CP1 (horizontal) e CP2 (vertical). Assim, o facto de as setas que servem de marcadores das variáveis correspondentes aos genes 1, 9 e 6 serem quase horizontais reflecte a existência de uma correlação forte positiva dessas variáveis com CP1. Ainda do biplot da Figura 3.2(a), uma forte correlação da CP2 com as variáveis gene 119, 306 e 878 é também ilustrada pelo facto de os marcadores dessas variáveis serem aproximadamente verticais.

Na leitura deste tipo de conclusões é necessário ter muita cautela, uma vez que a representação bidimensional é apenas uma aproximação. Como neste caso a proporção de variabilidade associada às duas primeiras CPs é 48.4%, as conclusões tiradas podem não corresponder à realidade. A título de curiosidade, e na tentativa de obter uma maior proporção de variabilidade para melhor aproximar os dados, realizámos um novo estudo, desta vez restringindo a matriz de dados aos 10 genes que apresentam maior variância.



(a) ACP Biplot sem recurso ao comando `biplot` para o conjunto Dados 1.

(b) ACP Biplot sem recurso ao comando `biplot` para o conjunto Dados 1 restrito aos 10 genes que apresentam maior variância.

Figura 3.2: Biplots obtidos através da ACP para o conjunto Dados 1.

Ordenámos as variáveis pelo desvio padrão:

```
> sort(sd(elementos))
```

e aplicámos ACP sobre a matriz de dados restrita. O sumário das proporções de variância explicadas pelas novas CP é o seguinte:

```
Importance of components:
                              PC1       PC2       PC3       PC4       PC5       PC6
Standard deviation     6138.120 4682.713 2.45e+03 2141.558 1.87e+03 1.49e+03
Proportion of Variance    0.476    0.277 7.61e-02    0.058 4.41e-02 2.81e-02
Cumulative Proportion     0.476    0.753 8.30e-01    0.888 9.32e-01 9.60e-01
```

Podemos verificar que, desta forma, 75.3% da variabilidade total dos dados restritos é preservada pela projecção da nuvem de pontos de $\mathbb{R}^{10}$ sobre o subespaço bi-dimensional, gerado pelas duas primeiras CPs. Analisando a nova representação gráfica (Figura 3.2(b)), verifica-se que se mantém o grau relativamente elevado de correlação entre os genes 878, 306 e 119 e que, neste caso, o gene 357 também está fortemente correlacionado com eles. A variância dos genes 878 e 306 continua a ser muito superior à dos outros genes. As setas que servem de marcadores das variáveis correspondentes aos genes 878, 306, 119 e 357 são praticamente horizontais, o que reflecte a forte correlação dessas variáveis com a CP1. A elevada correlação da CP2 com a variável gene 9 é também ilustrada pelo facto de o marcador dessa variável ser vertical. Além disso, nota-se que as variáveis gene 878, 306, 119 e

357 não estão correlacionadas com as variáveis gene 9 e 22, porque apresentam um ângulo de aproximadamente 90^{o}, o que reforça a ideia destas variáveis explicarem as novas componentes CP1 e CP2, respectivamente. Também é possível observar que os indivíduos T18, T29, T6, T34, T19, T17, T31, T32 e T30 são principalmente expressos à custa dos genes 1, 9, 6, 26 e 16. Neste caso, também se nota que existe um grande afastamento dos indivíduos T37 e T2 em relação aos outros, situação que já era visível na primeira análise (Figura 3.2(a)).

Retomando o conjunto Dados 1, sem restrições, aplicámos AC. Na AC, o número de eixos a serem seleccionados depende da percentagem acumulada das inércias principais e define-se a dimensionalidade de acordo com uma precisão considerada satisfatória e cuja visualização seja possível (uma, duas ou três dimensões). Para o conjunto Dados 1, do R fazendo

```
> summary(ca(elementos))
```

obtivemos

```
Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.042740  15.4  15.4  ************************
 2      0.031556  11.4  26.8  ******************
                                   ...
 60     0.000319   0.1  99.9
 61     0.000298   0.1 100.0
        -------- -----
 Total: 0.276766 100.0
```

Tal significa que em duas dimensões temos apenas uma precisão de 26.8%. Elaborámos o biplot correspondente à AC sobre os indivíduos sem (Figura 3.3(a)) e com (Figura 3.3(b)) massa. Por defeito, quando se realiza o gráfico para a AC, os indivíduos aparecem numerados pela ordem em que são lidos, pelo que identificámos na Tabela 3.2 a lista dos indivíduos e o número correspondente, para permitir a correcta leitura dos dados e interpretação dos resultados.

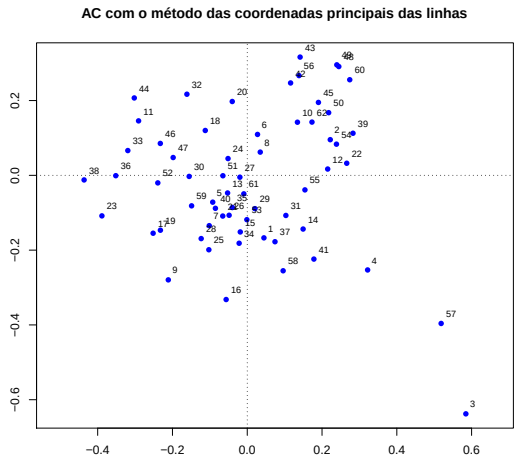
Para perceber a contribuição de cada ponto para a determinação dos eixos e o quanto cada ponto está representado em cada eixo, usa-se o que Greenacre (Greenacre [31]) denominou como sendo contribuição absoluta e contribuição relativa. A contribuição absoluta exprime o quanto um perfil contribui para a inércia principal do eixo em estudo. Graficamente, quanto maior for a intensidade da cor maior é a contribuição absoluta. A contribuição relativa exprime o quanto o perfil está representado neste eixo, e graficamente quanto maior for a intensidade da cor maior é a contribuição relativa (Nenadic and Greenacre [39]). Elaborámos

Individuos	Nº associado	Individuos	Nº associado	Individuos	Nº associado
T1	1	N11	22	N29	43
N1	2	T12	23	T29	44
T2	3	N12	24	T30	45
N2	4	T13	25	T31	46
T3	5	T14	26	T32	47
N3	6	T15	27	N32	48
T4	7	T16	28	T33	49
N4	8	T17	29	N33	50
T5	9	T18	30	N34	51
N5	10	T19	31	T34	52
T6	11	T20	32	T35	53
N6	12	T21	33	N35	54
T7	13	T22	34	N36	55
N7	14	T23	35	T36	56
T8	15	T24	36	T37	57
N8	16	T25	37	T38	58
T9	17	T26	38	T39	59
N9	18	N27	39	N39	60
T10	19	T27	40	T40	61
N10	20	T28	41	N40	62
T11	21	N28	42		

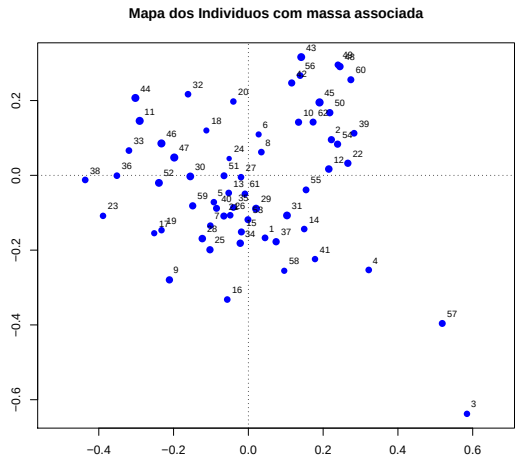
Tabela 3.2: Lista dos indivíduos com número correspondente utilizado na AC.

o biplot correspondente à AC sobre os indivíduos com contribuição absoluta (Figura 3.4(a)) e com contribuição relativa (Figura 3.4(b)).

A interpretação gráfica é feita em dois passos. Em primeiro lugar investigam-se os pontos em cada eixo separadamente. Da observação da Figura 3.4(a), verifica-se que a intensidade da cor é relativamente fraca, pelo que os pontos têm uma contribuição absoluta muito baixa. Nota-se que os perfis contribuem pouco para a inércia principal de ambos os eixos, o que já era de esperar devido à inércia explicada para o primeiro eixo ser de 15.4% e para o segundo 11.4%. Observando agora a Figura 3.4(b), os pontos que têm maior contribuição relativa correspondem na maioria aos indivíduos com células tumorais, e são: 56 (T36), 60 (N39), 3 (T2), 41 (T28), 53 (T35), 35 (T23), 40 (T27), 46 (T31), 33 (T21), 42 (N28), 45 (T30) e 4 (N2).

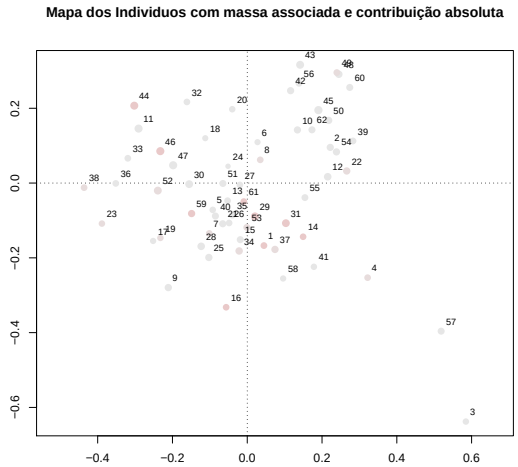


(a) AC biplot com recurso ao comando `plot(ca(···))` para as coordenadas principais dos indivíduos.

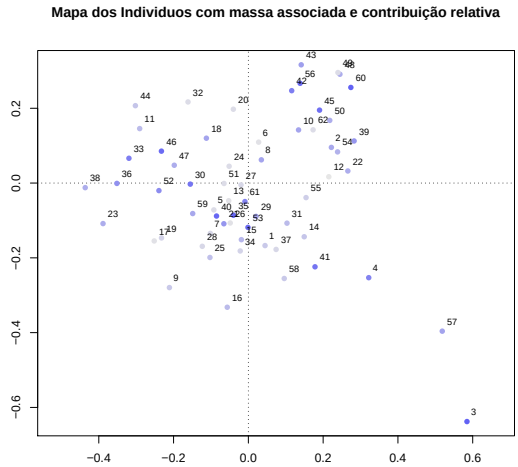


(b) AC biplot com recurso ao comando `plot(ca(···))` para as coordenadas principais dos indivíduos com massa.

Figura 3.3: Gráficos obtidos através da AC biplot para o conjunto Dados 1.



(a) AC biplot com recurso ao comando `plot(ca(···))` para as coordenadas principais dos indivíduos com contribuição absoluta.



(b) AC biplot com recurso ao comando `plot(ca(···))` para as coordenadas principais dos indivíduos com contribuição relativa.

Figura 3.4: Gráficos com as contribuições absolutas e relativas, obtidos através da AC biplot para o conjunto Dados 1.

Em segundo lugar, investiga-se a interpretação dos pontos no conjunto dos eixos. Da análise da Figura 3.3(b), resulta que os indivíduos com massa mais pequena estão, em geral,

agrupados e relativamente perto da origem. Evidencia-se um grupo no primeiro quadrante que corresponde ao grupo dos indivíduos com as células normais quando comparamos com a Figura 3.5, a qual corresponde ao AC biplot, conforme proposto em Nenadic and Greenacre [38]. Nesse biplot, os indivíduos com células normais aparecem, em geral, bem separados dos restantes indivíduos. Tal como já acontecia com a representação através da ACP, os indivíduos T2 e T37 estão muito afastados dos restantes. Em termos genéricos, podemos dizer que o eixo dos xx separa bem as células tumorais das normais (o que não acontecia com a representação obtida através da ACP), com excepção das células T2, T37, T36, T30, T33 e N8.

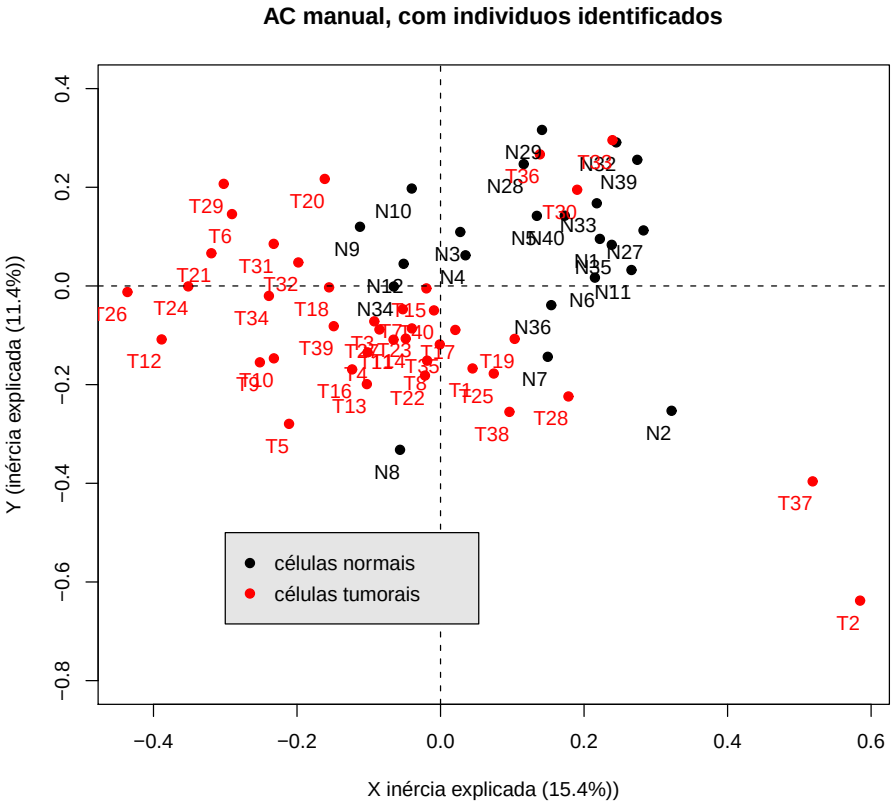
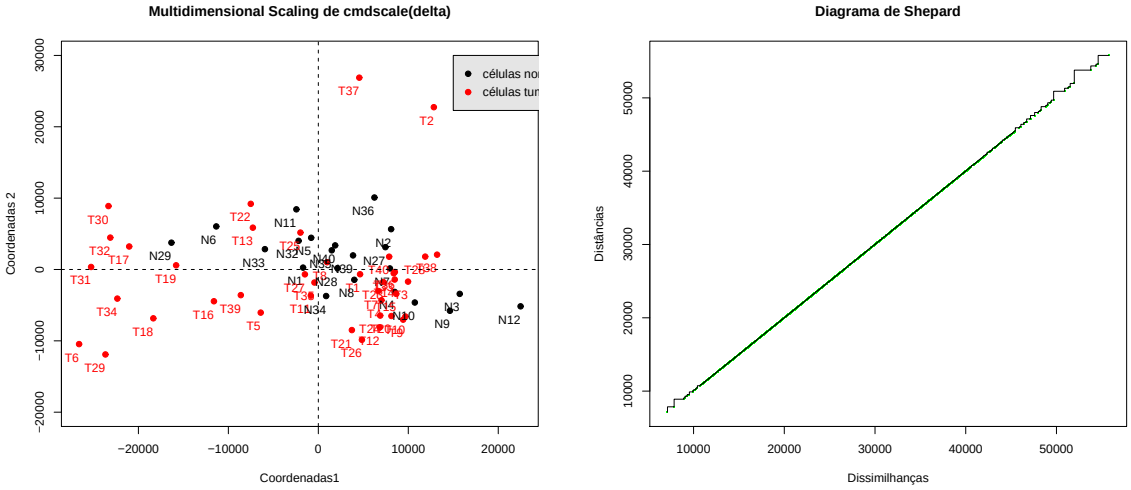


Figura 3.5: AC biplot sem recurso ao comando `plot(ca(···))` para as coordenadas principais dos indivíduos, com a identificação, para o conjunto Dados 1.

Aplicando MSD ao conjunto de Dados 1, conseguimos construir os gráficos apresentados na Figura 3.6. Com base na técnica MDS não obtivemos uma separação das células tumorais

das normais normais no conjunto de dados. Evidencia-se, tal como nos outros métodos, o afastamento dos indivíduos T37 e T2.



(a) MDS biplot com recurso ao comando `cmds-cale(...)` para o conjunto Dados 1.

(b) Diagrama de Shepard para o conjunto Dados 1.

Figura 3.6: Gráficos obtidos através do MDS biplot para o conjunto Dados 1.

A qualidade do escalonamento multidimensional pode ser avaliada graficamente através do diagrama de Shepard. Neste diagrama é confrontada as dissemelhanças originais com as distâncias obtidas a partir de um escalonamento multidimensional. A situação ideal é os pontos coincidirem com a bissectriz (Everitt and Hothorn [21]). No nosso caso, observando a Figura 3.6(b), podemos dizer que a aplicação da técnica MDS conduziu a uma representação no plano como se pretendia, ou seja, que mantém, aproximadamente, a distância euclideana original dos dados no espaço bidimensional.

3.2.2 Dados 2: Dados sobre pares de codões nas sequências de DNA

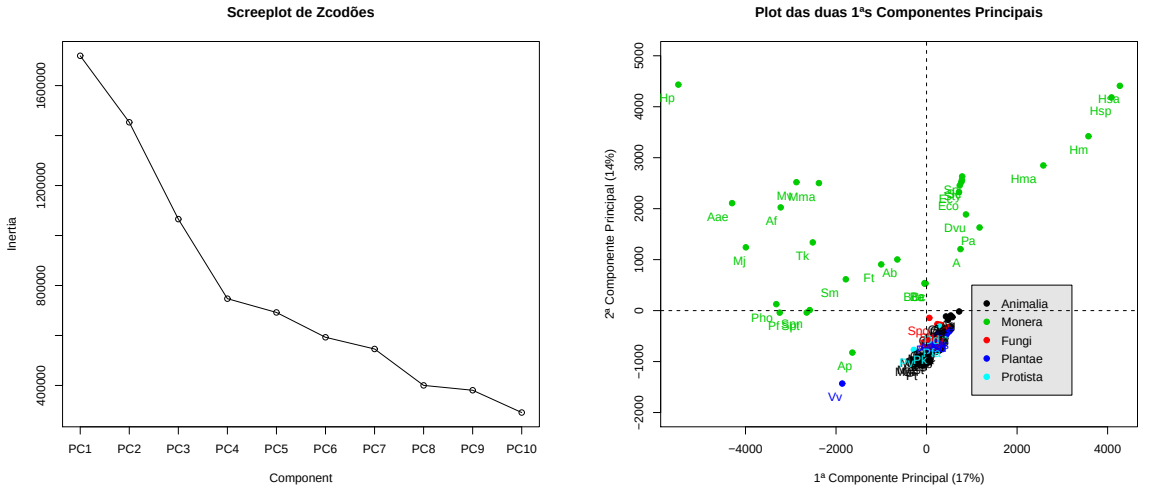
O conjunto Dados 2 contém um elevado número de variáveis o que torna difícil qualquer análise. Assim, propusemos aplicar as técnicas ACP, AC e MDS com o objectivo de reduzir o espaço das variáveis a um plano e construir biplots com vista a investigar regras associadas ao grau de preferência dos pares de codões no sequenciamento genómico das 123 espécies e/ou possíveis relações entre o grau de preferencia de pares de codões e o reino, ou género, a que as espécies pertencem.

À semelhança da análise dos Dados 1, começámos por aplicar ACP, depois AC (usando a *package* `ca`) e finalmente MDS. Os comandos do R utilizados para a ACP relativa aos Dados 2 encontram-se na Secção A.3.1, os utilizados para AC encontram-se na Secção A.3.3 e os

utilizados para o MDS encontram-se na Secção A.3.4.

Aplicando ACP, procedeu-se à análise das proporções de variância explicada pelas novas variáveis (CP). Do R obtivemos:

```
Importance of components:
                            PC1      PC2      PC3     PC4      PC5      PC6
Standard deviation     1311.231 1205.490 1032.316 864.444 831.9849 769.8917
Proportion of Variance    0.166    0.140    0.103   0.072   0.0667   0.0571
Cumulative Proportion     0.166    0.306    0.408   0.480   0.5468   0.6039
                            PC7      PC8      PC9     PC10     PC11     PC12
Standard deviation     738.9101 632.6514 617.0959 540.0262 500.3883 492.7880
Proportion of Variance   0.0526   0.0386   0.0367   0.0281   0.0241   0.0234
Cumulative Proportion    0.6565   0.6950   0.7317   0.7598   0.7839   0.8073
                          ...
                           PC123
Standard deviation     5.82e-13
Proportion of Variance 0.00e+00
Cumulative Proportion  1.00e+00
```



(a) Screeplot das Componentes Principais para Dados 2.

(b) Nuvem de pontos no espaço definido pelas duas primeiras CPs.

Figura 3.7: Gráficos obtidos através da ACP para o conjunto Dados 2.

Analisando a linha referente à proporção acumulada, verifica-se que a primeira CP explica aproximadamente 16.6% da variabilidade dos dados. Já as duas primeiras componentes acumularam uma percentagem de aproximadamente 30.6% da variabilidade dos dados. As demais

componentes absorvem cerca de 69.4% da variabilidade. Tendo em conta o screeplot (Figura 3.7(a)), podemos observar um primeiro cotovelo, um em "4" e outro em "8" sugerindo a escolha de 4 ou 8 componentes principais para a redução da dimensionalidade dos dados. Porém, prosseguiremos a análise tomando apenas as duas primeiras CPs em vez de quatro ou oito, com vista a obter uma visualização gráfica dos dados projectados num plano. Assim, teremos 30.6% da variabilidade total dos dados preservada na projecção da nuvem de pontos sobre o subespaço bi-dimensional de $\mathbb{R}^{123}$, gerado pelas duas primeiras CPs.

Apesar do valor 30.6% não corresponder a um valor elevado de variabilidade explicada, a observação da Figura 3.7(b) sugere que as duas primeiras CPs permitem diferenciar o comportamento das espécies do reino *Monera* das restantes espécies dos outros reinos.

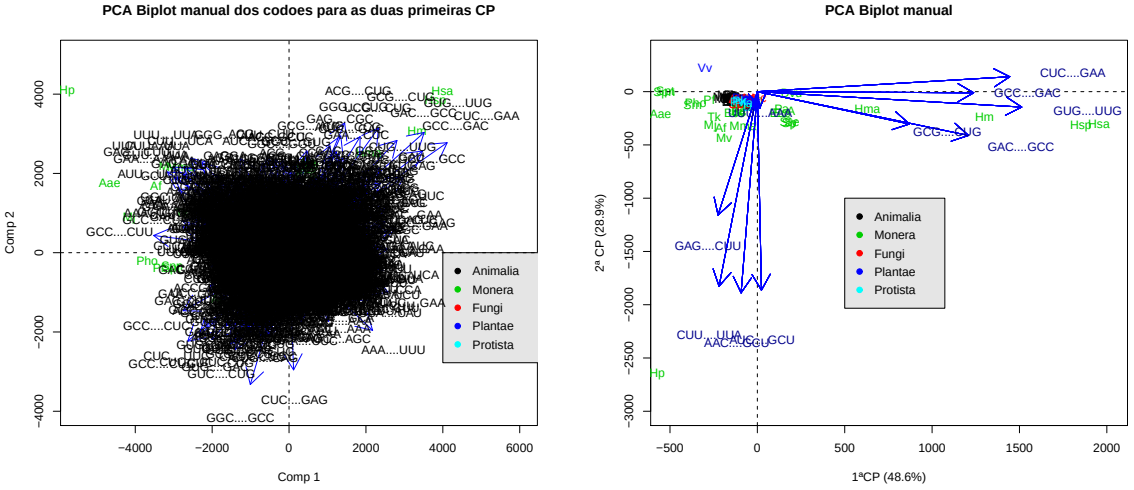
Elaborámos o biplot da ACP (ver Figura 3.8(a)). A representação simultânea das espécies e das variáveis num só gráfico impossibilita uma análise gráfica adequada dado o elevado número de variáveis (3721).

À semelhança da análise realizada para o conjunto Dados 1, procedemos a uma análise dos Dados 2 restinguindo o conjunto das váriáveis aos 10 pares de codões que apresentaram maior variância. Com o R obtivemos

```
Importance of components:
                          PC1     PC2      PC3      PC4     PC5     PC6
Standard deviation     314.582 242.450 125.8309 114.9306 82.7646 65.6230
Proportion of Variance   0.486   0.289   0.0778   0.0649  0.0337  0.0212
Cumulative Proportion    0.486   0.775   0.8531   0.9180  0.9517  0.9729
                          PC7      PC8      PC9     PC10
Standard deviation     60.0937 28.01842 24.21408 23.23681
Proportion of Variance  0.0177  0.00386  0.00288  0.00265
Cumulative Proportion   0.9906  0.99446  0.99735  1.00000
```

Desta forma, 77.5% da variabilidade total, do conjunto Dados 2 restrito àquelas 10 variáveis, é preservada pela projecção da nuvem de pontos sobre o subespaço bi-dimensional de $\mathbb{R}^{10}$ gerado pelas duas primeiras CPs. Analisando a nova representação gráfica, ilustrada na Figura 3.8(b) e relativa ao conjunto Dados 2 restrito, verificámos uma correlação elevada entre os pares de codões GCG-CUG e GAC-GCC . Observámos também que a variância do par UUU-AAA é muito inferior comparativamente com os restantes 9 pares de codões em análise. As setas que servem de marcadores das variáveis correspondentes ao pares de codões CUC-GAA, GCC-GAC, GCG-CUG, GAC-GCC e GUG-UUG são quase horizontais, o que reflecte uma correlação relativamente forte dessas variáveis com a CP1. Similarmente, a existência de correlação relativamente forte entre CP2 e as variáveis UUU-AAA, GAG-CUU,

CUU-UUA, AAC-GCU e AUC-GCU, tendo em conta que os marcadores dessas variáveis se apresentam em posição aproximadamente vertical. Salienta-se também, da Figura 3.8(b), que as espécies do reino *Monera* se encontram diferenciadas das restantes podendo ser divididas em dois grupos. Tal como na análise anterior, a espécie Hp está muito distante das restantes e está directamente correlacionada com a CP2. Nota-se também que a espécie Vv está distanciada das restantes plantas.



(a) ACP Biplot sem recurso ao comando `biplot` para o conjunto Dados 2.

(b) ACP Biplot sem recurso ao comando `biplot` para o conjunto Dados 2 restrito aos 10 pares de codões que apresentam maior variância.

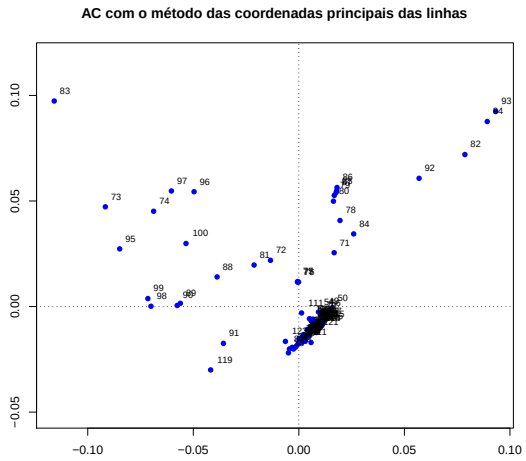
Figura 3.8: Biplots obtidos através da ACP para o conjunto Dados 2.

Aplicando AC sobre o conjunto total Dados 2, resulta que a percentagem acumulada das inércias principais apresenta uma precisão de 30.54%, como se deduz com auxilio do R.

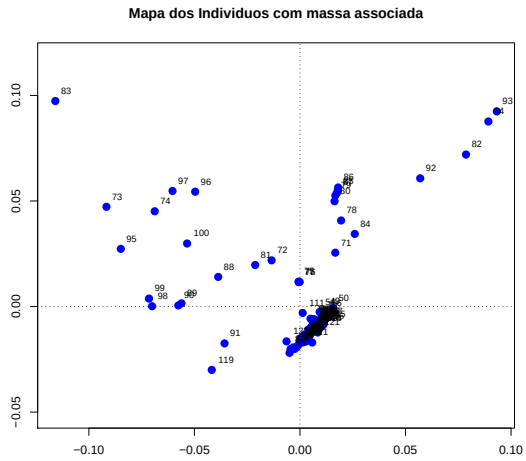
```
>print(head(summary(ca(novo))))

$scree
                values      values2   values3
 [1,]   1 7.935222e-04 1.658840e+01  16.58840
 [2,]   2 6.672874e-04 1.394948e+01  30.53788
 [3,]   3 4.909664e-04 1.026354e+01  40.80142
                                ...
```

Elaborámos o biplot correspondente à AC sobre as 123 espécies em análise sem (Figura 3.9(a)) e com (Figura 3.9(b)) massa.

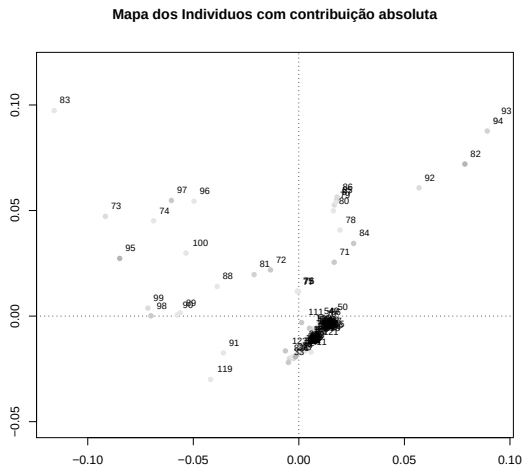


(a) AC biplot com recurso ao comando `plot(ca(···))` para as coordenadas principais das 123 espécies em análise.

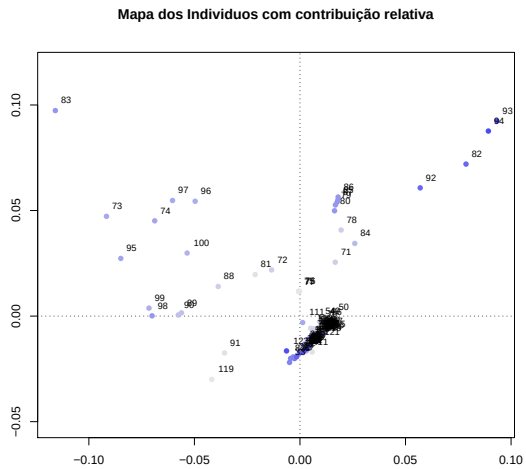


(b) AAC biplot com recurso ao comando `plot(ca(···))` para as coordenadas principais das 123 espécies em análise com massa.

Figura 3.9: Gráficos obtidos através da AC biplot para o conjunto Dados 2.



(a) AC biplot com recurso ao comando `plot(ca(···))` para as coordenadas principais das 123 espécies em análise com contribuição absoluta.



(b) AC biplot com recurso ao comando `plot(ca(···))` para as coordenadas principais das 123 espécies em análise com contribuição relativa.

Figura 3.10: Gráficos com as contribuições absolutas e relativas, obtidos através da AC biplot para o conjunto Dados 2.

Da análise da Figura 3.9(b) resulta que todas as espécies têm massa alta e aproximadamente igual. O tamanho do marcador indica a proporção da frequência relativa de cada ponto. Em geral, o gráfico mostra que as espécies têm frequências relativas semelhantes, destacando um grupo de espécies em particular. Analisando a Figura 3.11, contendo o AC biplot e onde as diferentes espécies são diferenciadas por cor, vemos que esse grupo corresponde aos reinos *Animalia*, *Fungi*, *Plantae* e *Protista*.

Da observação da Figura 3.10(a), verifica-se que a intensidade da cor é muito fraca, pelo que os pontos têm uma contribuição absoluta baixa. Nota-se que os perfis contribuem pouco para a inércia principal de ambos os eixos, o que já era de esperar devido à inércia explicada para o primeiro eixo ser de 16.59% e para o segundo 13.95%.

Analisando agora a Figura 3.10(b) notamos que os pontos que têm maior contribuição relativa estão muito dispersos e "espalhados" pelos diversos reinos.

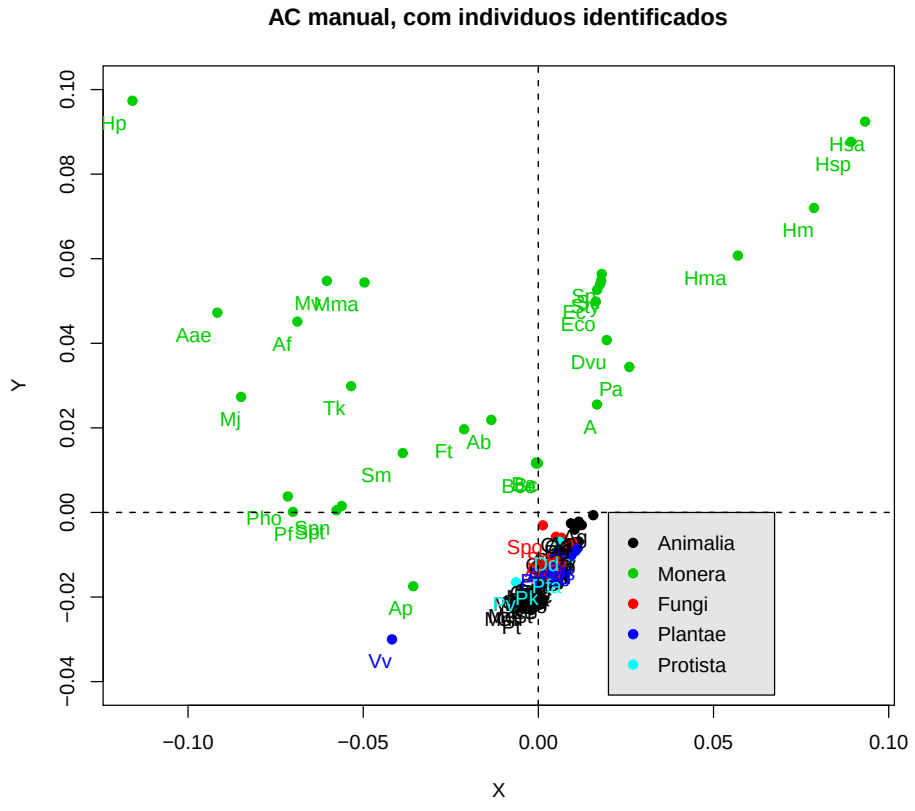
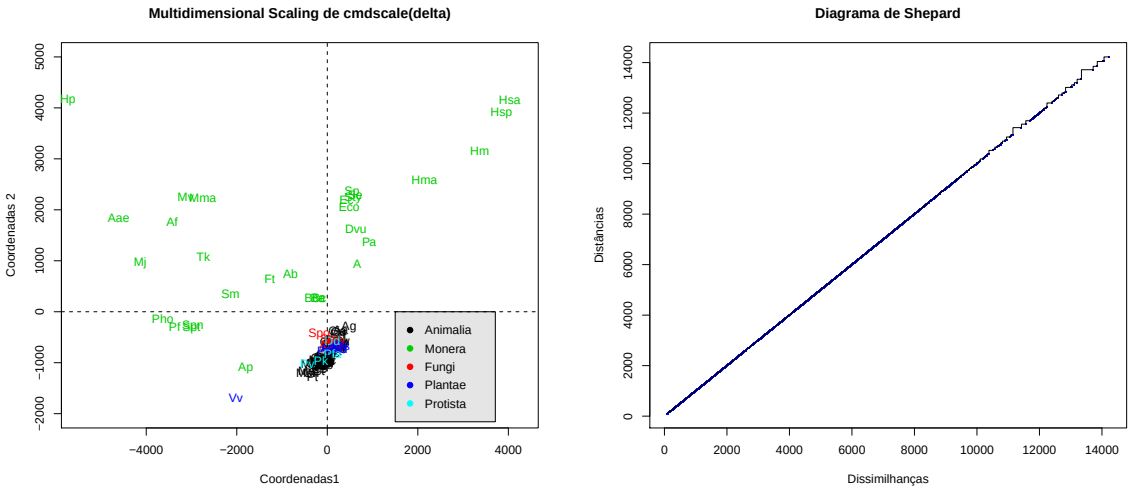


Figura 3.11: AC biplot sem recurso ao comando `plot(ca(···))` para as coordenadas principais das 123 espécies em análise, com a identificação, para o conjunto Dados 2.

Procedendo agora à obtenção manual dos gráficos, conforme feito por Greenacre e Nenadic (Nenadic and Greenacre [38]), obtivemos o gráfico (3.11). Da sua análise concluímos que o marcador correspondente à espécie Spo está relativamente perto da origem, o que implica que pouco contribui para o valor da estatística χ^2. O "mapa" obtido é em tudo semelhante ao que foi obtido através da ACP. Genericamente, o eixo dos xx separa bem o reino *Monera* dos restantes 4 reinos. Contudo, a espécie Vv, que pertence ao reino *Plantae*, está muito distante das outras e a espécie Ap está mais próxima da espécie Vv do que as espécies pertencentes ao seu reino.

Aplicando MSD ao conjunto de Dados 2, conseguimos construir os gráficos apresentados na Figura 3.12.

A Figura 3.12(a) é semelhante às que obtivemos através da ACP e da AC, destacando-se o reino *Monera* dos restantes reinos. No entanto, dentro desse reino existem muitas dissemelhanças. Na realidade, se analisarmos a nuvem de pontos desse reino, existem diversos subgrupos. As espécies que estão mais próximas, ou seja, mais semelhantes, correspondem, em geral, a espécies que pertencem ao mesmo género[3]. Assim, esta metodologia de MDS (mais precisamente, a Análise de Coordenadas Principais), forneceu de facto semelhanças/dissemelhanças concordantes com o esperado em termos biológicos. Avaliando a qualidade do escalonamento



(a) MDS biplot com recurso ao comando `cmds-cale(...)` para o conjunto Dados 2.

(b) Diagrama de Shepard para os Dados 2.

Figura 3.12: Gráficos obtidos através do MDS biplot para o conjunto Dados 2.

multidimensional através do diagrama de Shepard (Figura 3.12(b)), observa-se uma quase

[3]O género é identificável através do primeiro nome da espécie.

sobreposição dos valores das dissemelhanças originais com as distâncias obtidas a partir de um escalonamento multidimensional. Concluímos então que a técnica MDS sobre os Dados 2 produz uma representação das espécies num espaço bidimensional mantendo aproximadamente as distâncias euclideanas observadas no espaço original definido pelas 3721 variáveis.

Capítulo 4

Conclusões e Trabalhos Futuros

Ao longo do curso, estamos habituados a aplicar os conhecimentos adquiridos a exemplos académicos. Aplicá-los a dados reais constitui uma tarefa difícil mas ambiciosa.

Na presente dissertação três metodologias de redução da dimensionalidade de dados multidimensionais (ACP, ACP e MDS) e a correspondente representação dos dados em biplots foram estudados e aplicadas a três bases de dados. Considerámos a base de dados da *Iris*, amplamente divulgada na literatura especializada em Estatística Multivariada, e duas bases de dados mais recentes da Biologia Molecular, aqui denominadas Dados 1 e Dados 2.

A base Dados 1 diz respeito a uma experiência de *microarrays*, com resultados publicados pela primeira vez em 1999, e é constituída pelos níveis de expressão genética de 2000 genes (variáveis) sobre 62 tecidos amostrais distintos (indivíduos). A base Dados 2 resulta do cálculo de uma medida de associação, definida em termos da estatística de Pearson, entre dois codões. Esta base de dados não se encontra publicada e tem como propósito quantificar o grau de preferência dos 3721 possíveis pares de codões consecutivos (variáveis) contidos nas sequências de DNA de 123 espécies (indivíduos) divididas por 5 reinos diferentes. Em ambas as bases, Dados 1 e Dados 2, o número de variáveis é, comparativamente, bastante superior ao número de indivíduos.

Ao contrário daquelas duas, a base de dados *Iris* contém um número de variáveis (4) inferior ao número de indivíduos (150). Dado o carácter, em geral, académico associado à base *Iris*, esta foi aqui utilizada para testar e ilustrar as saídas dos programas que implementámos no R com o objectivo de construir biplots com maior flexibilização nas etiquetas e coloração dos pontos nos gráficos.

Com os Dados 1 comparámos os nossos resultados com os obtidos em Park et al. [44]. Com os Dados 2, tentámos extrair informação a nível de possíveis agrupamentos de reinos.

Para todas as bases de dados começámos por aplicar ACP, a seguir AC e por último MDS. Sobre cada metodologia construímos os correspondentes biplots.

Aplicando ACP, para ambas as bases de dados da Biologia Molecular aqui consideradas, a variância total explicada não foi elevada: 48.8% para Dados 1 e 30.6% para Dados 2. Tal facto advém do elevado número de variáveis em estudo: 2000 em Dados 1 e 3721 em Dados 2.

Por inspecção visual sobre os biplots verificámos que, usando a técnica de ACP sobre Dados 1, não nos foi possível estabelecer uma separação das células normais das tumorais com base nas duas primeiras CPs [1]. Aplicando ACP sobre o conjunto de Dados 2 foi possível obter uma separação das espécies do reino Monera das restantes. Porém, o elevado número de variáveis envolvidos não possibilitou a detecção de uma regra na preferência dos pares de codões que diferencie aqueles dois grupos.

Aplicando AC, o método conduziu a uma separação das células tumorais das normais num espaço bidimensional para os Dados 1. Para a base Dados 2 obtivemos a mesma representação que foi obtida usando ACP.

Aplicando MDS, para a base Dados 1 não conseguimos obter uma separação das células tumorais das normais e, para a base Dados 2, obtivemos a mesma representação à obtida com as outras duas metodologias.

Como trabalho futuro propomos explorar a base Dados 2 restringindo as variáveis a conjuntos de pares de codões com propriedades definidas sob pontos de vista biológicos, como por exemplo, pares contendo os nucleótidos G e C versus pares não contendo os nucleótidos C e G ou, outro exemplo, pares de codões que codificam os mesmos pares de aminoácidos.

Futuramente, seria também interessante realizar um estudo mais detalhado tomando as espécies do reino *Animalia*, *Plantae*, *Fungi* e *Protista* uma vez que são aquelas que todos os biplots mostram menos informação em termos de relações eventualmente existentes nos dados.

[1] Este resultado contrariava o trabalho apresentado em Park et al. [43]: o biplot correspondente (Figura 3.1(b)) não coincidia com o biplot apresentado na Figura 4a daquele artigo como seria esperado. Contactados os autores por email fomos informados da existência de uma gralha sendo que Figura 4a apresentada foi obtida através da AC e não da ACP; correctamente deveria estar um gráfico igual ao da Figura 3.1(b).

Apêndice A

Comandos usados no R

A.1 Dados *Iris*

```
> library(rJava);library(xlsxjars);library(xlsx)
> library(graphics); library(stats); library(MASS)
> dados <- read.xlsx("plantas.xlsx", 1)
> caracteristicas<- data.frame(dados)
> attach(caracteristicas)
> iris <- dados[,1:4]
```

A.1.1 ACP para os dados *Iris*

Para desenvolver os cálculos necessários desta secção foram consultadas as seguintes referências bibliográficas: (hel [8]), (hel [6]), (hel [7]), (hel [2]), (Etienne [20]), (hel [4]) e (hel [5]).

```
> iris.pc<-prcomp(iris,center=TRUE,scale=FALSE)
> iris.pc
> summary(iris.pc)
> iris.pc$x
> screeplot(iris.pc,type="l",main="Screeplot de iris.pc")
> plot(iris.pc$x[,1:2], col=rep(especies),
+ main="Plot das duas 1ªs Componentes Principais",
+xlab="1ª Componente Principal",
+ ylab="2ª Componente Principal", type="p", pch=19)
> legend(0,-0.8,c("Iris setosa","Iris versicolor","Iris virginica"),
+ col=c(1,2,3), text.col="black",bg='gray90',pch=19)
> abline(h=0,lty=2); abline(v=0,lty=2)
> var(iris)
> cor(iris,prcomp(iris)$x)
```

```
> cor(iris)
> biplot(iris.pc,choices=1:2,main="Biplot das 2 1ªs CP",xlab="1ª CP",
+ ylab="2ª CP",
+ var.axes=TRUE,scale=0,pc.biplot=TRUE)
> abline(h=0,lty=2); abline(v=0,lty=2)
> N<-nrow(iris)
> setosa<-rep(1,50); versicolor<-rep(2,50); virginica<-rep(3,50)
> grupos<-c(setosa,versicolor,virginica)
> acp<-prcomp(Z,scale=FALSE, center=TRUE)
> xmin <- min(acp$x[,1])
> xmax <- max(acp$x[,1])
> ymin <- min(acp$x[,2])
> ymax <- max(acp$x[,2])
> plot(c(xmin,xmax),c(ymin,ymax),col="white",
+ xlab="1ª Componente Principal",
+  ylab="2ª Componente Principal")
> text(iris.pc$x[,1],iris.pc$x[,2],1:N,col=grupos)
> title("PCA Biplot manual das iris para as duas primeiras CP")
> legend(2,1.2,c("Iris setosa","Iris versicolor","Iris virginica"),
+ col=c(1,2,3), text.col="black",bg='gray90',pch=19)
> abline(v=0,lty=2); abline(h=0, lty=2)
> xl.min = min(0,min(acp$rotation[,1]))
> xl.max = max(0,max(acp$rotation[,1]))
> yl.min = min(0,min(acp$rotation[,2]))
> yl.max = max(0,max(acp$rotation[,2]))
> xl.scale = max(abs(xmax),abs(xmin))/max(abs(xl.max),abs(xl.min))*0.75
> yl.scale = max(abs(ymax),abs(ymin))/max(abs(yl.max),abs(yl.min))*0.75
> arrows(rep(0,100),rep(0,100),acp$rotation[,1]*xl.scale,
+ acp$rotation[,2]*yl.scale, col="blue")
> text(acp$rotation[,1]*xl.scale, acp$rotation[,2]*yl.scale ,
+ colnames(iris))
```

A.1.2 AC para os dados *Iris*

Para desenvolver os cálculos necessários desta secção foram consultadas as seguintes referências bibliográficas: (Nenadic and Greenacre [40]), (Nenadic and Greenacre [39]) e (Nenadic and Greenacre [38]).

```
> library(vegan); library(rgl) #para poder utilizar a package ca
> library(ca)
> ca.novo <- ca(iris)
> summary(ca(iris))
```

```
> plot(ca(elementos), map="rowprincipal", what=c("all","none"),
+ mass=c(TRUE,FALSE),labels=c(2,0)
+ ,main="Mapa dos Individuos com massa associada")
> P<- iris/sum(iris)
> rm<-apply(P,1,sum)
> cm<-apply(P,2,sum)
> Q<-P-rm%*%t(cm)
> Dr<-as.matrix(diag(sqrt(1/rm)))
> Dc<-as.matrix(diag(sqrt(1/cm)))
> S<-Q/sqrt(rm%*%t(cm))
> s<-svd(S)
> U<-s$u
> V<-s$v
> Sigma<-diag(s$d)
> F<-Dr %*% U %*% Sigma
> ca.novo.raw<-ca.novo
> ca.novo.raw$rowcoord<-F
> x<-ca.novo.raw$rowcoord[,1]
> y<-ca.novo.raw$rowcoord[,2]
> plot(x,y,pch=19,col=rep(especies),xlab="X",
+ ylab="Y",ylim=c(-0.15,0.15))
> title("AC manual")
> text(x,y,lab=labels(row.names(iris)),1:N,col=grupos)
> abline(v=0,lty=2); abline(h=0, lty=2)
> legend(-0.3,.10,c("Iris setosa","Iris versicolor","Iris virginica"),
+ col=c(1,2,3), text.col="black",bg='gray90',pch=19)
```

A.1.3 MDS para os dados *Iris*

Para desenvolver os cálculos necessários desta secção foram consultadas as seguintes referências bibliográficas: (hel [3]) e (Everitt and Hothorn [21]).

```
> delta<-as.matrix(dist(iris),150,150)
> delta.mds<- cmdscale(delta, k=149, eig=TRUE)
> delta.mds<- cmdscale(delta, k=149, eig=TRUE, add=TRUE)
> x<-delta.mds$points[,1]
> y<-delta.mds$points[,2]
> plot(x, y, xlab="Coordenadas1",
+ ylab="Coordenadas 2",pch=19,col=rep(especies),ylim=c(-3,3))
> title("Multidimensional Scaling de cmdscale(delta)")
> legend(0,2,c("Iris setosa","Iris versicolor","Iris virginica"),
+ col=c(1,2,3), text.col="black",bg='gray90',pch=19)
```

```
> abline(v=0,lty=2); abline(h=0, lty=2)
> #Diagrama de Shepard
> deltaSh<-Shepard(delta[lower.tri(delta)],delta.mds$points)
> plot(deltaSh,pch=".",xlab="Dissimilhanças",
+ ylab="Distâncias", xlim= range(deltaSh$x),
+ ylim=range(deltaSh$x),col="red")
> title("Diagrama de Shepard")
> lines(deltaSh$x,deltaSh$yf,type="S")
```

A.2 Dados 1: Dados de *microarrays*

```
> library(rJava); library(xlsxjars); library(xlsx)
> library(graphics); library(stats); library(MASS)
> dados <- read.xlsx("matriz.xlsx", 1) #dados completos
> caracteristicas<- data.frame(dados)
> attach(caracteristicas)
> elementos <- dados[,2:2001] #dados completos
```

A.2.1 ACP para os Dados 1

Para desenvolver os cálculos necessários desta secção foram consultadas as seguintes referências bibliográficas: (hel [8]), (hel [6]), (hel [7]), (hel [2]), (Etienne [20]), (hel [4]) e (hel [5]).

```
> elementos.pc<-prcomp(elementos,scale=FALSE, center=TRUE)
> summary(elementos.pc)
> screeplot(elementos.pc,type="l",main="Screeplot elementos.pc")
> N=nrow(elementos)
> grupos<-c(2,1,2,1,2,1,2,
+ 1,2,1,2,1,2,1,2,1,
+ 2,1,2,1,2,1,2,1,2,
+ 2,2,2,2,2,2,2,2,2,2,
+ 2,2,2,1,2,2,1,1,2,2,2,
+ 2,1,2,1,1,2,2,1,1,2,2,
+ 2,2,1,2,1)
> plot(elementos.pc$x[,1:2],col=rep(Tipo),main="PCA Plot das 2 1ªs CP",
+xlab="1ª CP", ylab="2ª CP", type="p", pch=19)
> legend(10000,-15000,c("células normais","células tumorais"),
+ col=c(1,2), text.col="black",bg='gray90',pch=19)
> abline(h=0,lty=2); abline(v=0,lty=2)
> biplot(elementos.pc,choices=1:2,main="Biplot das 2 1ªs CP",
+xlab="1ª CP", ylab="2ª CP",
```

```
+ var.axes=FALSE,scale=1,pc.biplot=TRUE)
> abline(h=0,lty=2); abline(v=0,lty=2)
> #     Fazer o PCA Biplot Manualmente
> acp=prcomp(elementos,cor=FALSE,scale=FALSE, center=TRUE)
> xmin = min(acp$x[,1])
> xmax = max(acp$x[,1])
> ymin = min(acp$x[,2])
> ymax = max(acp$x[,2])
> plot(c(xmin,xmax),c(-30000,ymax),col="white", xlab="1ªCP (36.1%)",
+ ylab="2ª CP (12.3%)")
> text(acp$x[,1],acp$x[,2],1:N,col= grupos,labels=Ind)
> title("PCA Biplot manual")
> legend(10000,-20000,c("células normais","células tumorais"),
+ col=c(1,2), text.col="black",bg='gray90',pch=19)
> abline(v=0,lty=2); abline(h=0, lty=2)
> xl.min = min(0,min(acp$rotation[,1]))
> xl.max = max(0,max(acp$rotation[,1]))
> yl.min = min(0,min(acp$rotation[,2]))
> yl.max = max(0,max(acp$rotation[,2]))
> xl.scale = max(abs(xmax),abs(xmin))/max(abs(xl.max),abs(xl.min))*0.75
> yl.scale = max(abs(ymax),abs(ymin))/max(abs(yl.max),abs(yl.min))*0.75
> arrows(rep(0,100),rep(0,100),acp$rotation[,1]*xl.scale,
+acp$rotation[,2]*yl.scale,col="blue")
> text(acp$rotation[,1]*xl.scale*1.25,
+acp$rotation[,2]*yl.scale*1.25 ,
+ colnames(elementos),col="darkblue")
```

A.2.2 ACP para o conjunto Dados 1 restrito

```
> library(rJava); library(xlsxjars); library(xlsx)
> library(graphics); library(stats); library(MASS)
> dadosVar <- read.xlsx("matriz.xlsx", 4)
> caracteristicasVar<- data.frame(dadosVar)
> attach(caracteristicasVar)
> el <- dadosVar[,2:11]
> DcentEl<-scale(el,center=TRUE,scale=FALSE)
> s<-svd(DcentEl)
> Sigma<-diag(s$d)
> U<- s$u
> V<-s$v
> G<-U%*%Sigma
```

```
> H<-V
> Z<-G%*%t(H)
> el.pc<-prcomp(Z,scale=FALSE, center=TRUE)
> el.pc
> summary(el.pc)
> screeplot(el.pc,type="l",main="Screeplot de Z")
> N=nrow(elementos)
> grupos<-c(2,1,2,1,2,1,2,
+ 1,2,1,2,1,2,1,2,1,
+ 2,1,2,1,2,1,2,1,2,
+ 2,2,2,2,2,2,2,2,2,2,
+ 2,2,2,1,2,2,1,1,2,2,2,
+ 2,1,2,1,1,2,2,1,1,2,2,
+ 2,2,1,2,1)
> acp=prcomp(Z,cor=FALSE)
> xmin = min(acp$x[,1])
> xmax = max(acp$x[,1])
> ymin = min(acp$x[,2])
> ymax = max(acp$x[,2])
> plot(c(xmin,xmax),c(-20000,20000),col="white", xlab="1ª CP (47.6%)",
+ ylab="2ª CP (27.7%)")
> text(acp$x[,1],acp$x[,2],1:N,col= grupos,labels=Ind)
> title("PCA Biplot manual")
> legend(-15000,-10000,c("células normais","células tumorais"),
+ col=c(1,2), text.col="black",bg='gray90',pch=19)
> abline(v=0,lty=2); abline(h=0, lty=2)
> xl.min = min(0,min(acp$rotation[,1]))
> xl.max = max(0,max(acp$rotation[,1]))
> yl.min = min(0,min(acp$rotation[,2]))
> yl.max = max(0,max(acp$rotation[,2]))
> xl.scale = max(abs(xmax),abs(xmin))/max(abs(xl.max),abs(xl.min))*0.75
> yl.scale = max(abs(ymax),abs(ymin))/max(abs(yl.max),abs(yl.min))*0.75
> arrows(rep(0,100),rep(0,100),acp$rotation[,1]*xl.scale,
+  acp$rotation[,2]*yl.scale,col="blue")
> text(acp$rotation[,1]*xl.scale*1.25, acp$rotation[,2]*yl.scale*1.25 ,
+ colnames(el),col="darkblue")
```

A.2.3 AC para os Dados 1

Para desenvolver os cálculos necessários desta secção foram consultadas as seguintes referências bibliográficas: (Nenadic and Greenacre [40]), (Nenadic and Greenacre [39]) e (Nenadic

and Greenacre [38]).

```
> library(rgl) #para poder utilizar a package ca
> library(ca)
> ca.novo <- ca(elementos)
> #Figura 1 Mapa dos Individuos
> plot(ca(elementos), map="rowprincipal", what=c("all","none"),
+ mass=c(FALSE,FALSE),labels=c(2,2),arrows=c(FALSE,TRUE)
+ ,main="AC com o método das coordenadas principais das linhas")
> #Figura 2 Mapa dos Individuos com massa associada
> plot(ca(elementos), map="rowprincipal", what=c("all","none"),
+ mass=c(TRUE,FALSE),labels=c(2,0)
+ ,main="Mapa dos Individuos com massa associada")
> #Figura 3 Mapa dos Individuos com massa associada e contribuição absoluta
> plot(ca(elementos), map="rowprincipal", what=c("all","none"),
+ mass=c(TRUE,FALSE),labels=c(2,0),col=rep(Tipo),
+contrib=c("absolute","none")
+ ,main="Mapa dos Individuos com massa associada e contribuição absoluta")
> #Figura 4 Mapa dos Individuos com massa associada e contribuição relativa
> plot(ca(elementos), map="rowprincipal", what=c("all","none"),
+ mass=c(FALSE,FALSE),labels=c(2,0),contrib=c("relative","none")
+ ,main="Mapa dos Individuos com massa associada e contribuição relativa")
> inertia.rel <- sum(ca.novo$sv^2)
> # AC manual
> P<- elementos/sum(elementos)
> rm<-apply(P,1,sum)
> cm<-apply(P,2,sum)
> Q<-P-rm%*%t(cm)
> Dr<-as.matrix(diag(sqrt(1/rm)))
> Dc<-as.matrix(diag(sqrt(1/cm)))
> S<-Q/sqrt(rm%*%t(cm))
> s<-svd(S)
> U<-s$u
> V<-s$v
> Sigma<-diag(s$d)
> F<-Dr %*% U %*% Sigma
> ca.novo.raw$rowcoord<-F
> x<-ca.novo.raw$rowcoord[,1]
> y<-ca.novo.raw$rowcoord[,2]
> # Com legendas habituais
> plot(x,y,pch=19,col=rep(Tipo),xlab="X", ylim=c(-0.8,0.4),
+ ylab="Y")
```

```
> title("AC manual, com individuos identificados")
> text(x,y,labels=Ind,1:N,col=grupos)
> abline(v=0,lty=2); abline(h=0, lty=2)
> legend(-.3,-.5,c("células normais","células tumorais"),
+ col=c(1,2), text.col="black",bg='gray90',pch=19)
```

A.2.4 MDS para os Dados 1

Para desenvolver os cálculos necessários desta secção foram consultadas as seguintes referências bibliográficas: (hel [3]) e (Everitt and Hothorn [21]).

```
> delta<-as.matrix(dist(elementos),62,62)
> loc<-cmdscale(delta)
> x<--loc[,1]
> y<-loc[,2]
> plot(x, y, xlab="Coordenadas1",
+ ylab="Coordenadas 2",ylim=c(-20000,30000),pch=19,col=rep(Tipo))
> title("Multidimensional Scaling de cmdscale(delta)")
> text(x,y,labels=Ind,1:N,col=grupos)
> legend(15000,30000,c("células normais","células tumorais"),
+ col=c(1,2), text.col="black",bg='gray90',pch=19)
> abline(v=0,lty=2)
> abline(h=0, lty=2)
#Diagrama de Shepard
>deltaSh<-Shepard(delta[lower.tri(delta)],delta.mds$points)
>plot(deltaSh,pch=".",xlab="Dissimilhanças",
>ylab="Distâncias", xlim= range(deltaSh$x),
>ylim=range(deltaSh$x),col="green")
>title("Diagrama de Shepard")
>lines(deltaSh$x,deltaSh$yf,type="S")
```

A.3 Dados 2: Dados sobre pares de codões nas sequências de DNA

```
> library(rJava); library(xlsxjars); library(xlsx)
> library(graphics); library(stats); library(MASS)
> dados <- read.xlsx("experiencia3.xlsx", 1)
> caracteristicas<- data.frame(dados)
> attach(caracteristicas) #mostar as colunas de dados como variáveis
> X <- dados[,4:3724]
```

A.3.1 ACP para os Dados 2

Para desenvolver os cálculos necessários desta secção foram consultadas as seguintes referências bibliográficas: (hel [8]), (hel [6]), (hel [7]), (hel [2]), (Etienne [20]), (hel [4]) e (hel [5]).

```
> Dcent<-scale(X,center=TRUE,scale=FALSE)
> s<-svd(Dcent)
> Sigma<-diag(s$d)
> U<- s$u
> V<-s$v
> G<-U%*%Sigma
> H<-V
> Z<-G%*%t(H)
> codoes.pc<-prcomp(Z,scale=FALSE, center=TRUE)
> summary(codoes.pc)
> screeplot(codoes.pc,type="l",main="Screeplot de Zcodões")
> N<-nrow(X)
> animal<-rep(1,70)
> monera<-rep(3,30)
> fungi<-rep(2,11)
> plants<-rep(4,8)
> protozoarios<-rep(5,4)
> grupos<-c(animal,monera,fungi,plants,protozoarios)
> plot(codoes.pc$x[,1:2],
+ col=rep(Reino),
+ main="Plot das duas 1ªs Componentes Principais",
+ xlab="1ª Componente Principal (17%)",
+ ylab="2ª Componente Principal (14%)",
+ type="p", pch=19,ylim=c(-2000,5000))
> text(codoes.pc$x[,1],codoes.pc$x[,2],
+ 1:N,col=grupos,labels=Abreviatura )
> legend(1000,500,c("Animalia","Monera","Fungi","Plantae", "Protista"),
+ col=c(1,3,2,4,5), text.col="black",bg='gray90',pch=19)
> abline(h=0,lty=2); abline(v=0,lty=2)
> sort(sd(X))
> #     Fazer o PCA Biplot Manualmente
> acp=prcomp(Z,cor=FALSE)
> # Compute de min/max of new coordinates
> xmin = min(acp$x[,1])
> xmax = max(acp$x[,1])
> ymin = min(acp$x[,2])
```

```
> ymax = max(acp$x[,2])
> plot(c(xmin,6000),c(-4000,5000),col="white", xlab="Comp 1",
+ ylab="Comp 2")
> text(codoes.pc$x[,1],codoes.pc$x[,2],
+ 1:N,col=grupos,labels=Abreviatura )
> title("PCA Biplot manual dos codoes para as duas primeiras CP")
> legend(1700,0,c("Animalia","Monera","Fungi","Plantae", "Protista"),
+ col=c(1,3,2,4,5), text.col="black",bg='gray90',pch=19)
> abline(h=0,lty=2); abline(v=0,lty=2)
> xl.min = min(0,min(acp$rotation[,1]))
> xl.max = max(0,max(acp$rotation[,1]))
> yl.min = min(0,min(acp$rotation[,2]))
> yl.max = max(0,max(acp$rotation[,2]))
> xl.scale = max(abs(xmax),abs(xmin))/max(abs(xl.max),abs(xl.min))*0.75
> yl.scale = max(abs(ymax),abs(ymin))/max(abs(yl.max),abs(yl.min))*0.75
> arrows(rep(0,100),rep(0,100),acp$rotation[,1]*xl.scale,
+acp$rotation[,2]*yl.scale, col="blue")
> text(acp$rotation[,1]*xl.scale*1.25,
+ acp$rotation[,2]*yl.scale*1.25 ,colnames(X))
```

A.3.2 ACP para o conjunto de Dados 2 restrito

```
> library(rJava); library(xlsxjars); library(xlsx)
> library(graphics); library(stats);library(MASS)
> dadosVar <- read.xlsx("experiencia.xlsx", 1)
> caracteristicasVar<- data.frame(dadosVar)
> attach(caracteristicasVar)
> el <- dadosVar[,4:13]
> DcentEl<-scale(el,center=TRUE,scale=FALSE)
> s<-svd(DcentEl)
> Sigma<-diag(s$d)
> U<- s$u
> V<-s$v
> G<-U%*%Sigma
> H<-V
> Z<-G%*%t(H)
> el.pc<-prcomp(Z,scale=FALSE, center=TRUE)
> el.pc
> summary(el.pc)
> screeplot(el.pc,type="l",main="Screeplot de Z")
> N<-nrow(el)
```

```
> animal<-rep(1,70)
> monera<-rep(3,30)
> fungi<-rep(2,11)
> plants<-rep(4,8)
> protozoarios<-rep(5,4)
> grupos<-c(animal,monera,fungi,plants,protozoarios)
> acp=prcomp(Z,cor=FALSE)
> xmin = min(acp$x[,1])
> xmax = max(acp$x[,1])
> ymin = min(acp$x[,2])
> ymax = max(acp$x[,2])
> plot(c(xmin,xmax),c(-3000,ymax),col="white", xlab="1ªCP (48.6%)",
+ ylab="2ª CP (28.9%)")
> text(acp$x[,1],acp$x[,2],1:N,col= grupos,labels=Abreviatura)
> title("PCA Biplot manual")
> legend(500,-1000,c("Animalia","Monera","Fungi","Plantae", "Protista"),
+ col=c(1,3,2,4,5), text.col="black",bg='gray90',pch=19)
> abline(h=0,lty=2); abline(v=0,lty=2)
> xl.min = min(0,min(acp$rotation[,1]))
> xl.max = max(0,max(acp$rotation[,1]))
> yl.min = min(0,min(acp$rotation[,2]))
> yl.max = max(0,max(acp$rotation[,2]))
> xl.scale = max(abs(xmax),abs(xmin))/max(abs(xl.max),abs(xl.min))*0.75
> yl.scale = max(abs(ymax),abs(ymin))/max(abs(yl.max),abs(yl.min))*0.75
> arrows(rep(0,100),rep(0,100),acp$rotation[,1]*xl.scale,
+  acp$rotation[,2]*yl.scale,col="blue")
> text(acp$rotation[,1]*xl.scale*1.25, acp$rotation[,2]*yl.scale*1.25 ,
+ colnames(el),col="darkblue")
```

A.3.3 AC para os Dados 2

Para desenvolver os cálculos necessários desta secção foram consultadas as seguintes referências bibliográficas: (Nenadic and Greenacre [40]), (Nenadic and Greenacre [39]) e (Nenadic and Greenacre [38]).

```
> library(rgl) #para poder utilizar a package ca
> library(ca)
> novo<-X+abs(min(X)) #shifting Xij
> ca.novo <- ca(novo)
> #Figura 1 Mapa dos Individuos
> plot(ca(novo), map="rowprincipal", what=c("all","none"),
+ mass=c(FALSE,FALSE),labels=c(2,0),arrows=c(FALSE,TRUE)
```

```
+ ,main="AC com o método das coordenadas principais das linhas")
#Figura 2 Mapa dos Individuos com massa associada
plot(ca(novo), map="rowprincipal", what=c("all","none"),
mass=c(TRUE,FALSE),labels=c(2,0)
,main="Mapa dos Individuos com massa associada")
>#Figura 3 Mapa dos Individuos com contribuição absoluta
>plot(ca(novo), map="rowprincipal", what=c("all","none"),
+mass=c(FALSE,FALSE),labels=c(2,0),col=rep(Reino),contrib=c("absolute","none")
+,main="Mapa dos Individuos com contribuição absoluta")
>#Figura 4 Mapa dos Individuos com contribuição relativa
>plot(ca(novo), map="rowprincipal", what=c("all","none"),
+mass=c(FALSE,FALSE),labels=c(2,0),contrib=c("relative","none")
+,main="Mapa dos Individuos com contribuição relativa")
> inertia.rel <- sum(ca.novo$sv^2)
> # AC manual
> P<- novo/sum(novo)
> rm<-apply(P,1,sum)
> cm<-apply(P,2,sum)
> Q<-P-rm%*%t(cm)
> Dr<-as.matrix(diag(sqrt(1/rm)))
> Dc<-as.matrix(diag(sqrt(1/cm)))
> S<-Q/sqrt(rm%*%t(cm))
> s<-svd(S)
> U<-s$u
> V<-s$v
> Sigma<-diag(s$d)
> F<-Dr %*% U %*% Sigma
> ca.novo.raw<-ca.novo
> ca.novo.raw$rowcoord<-F
> x<-ca.novo.raw$rowcoord[,1]
> y<-ca.novo.raw$rowcoord[,2]
> # Com legendas habituais
> plot(x,y,pch=19,col=rep(Reino),xlab="X",ylim=c(-0.04,0.10),
+ ylab="Y")
> title("AC manual, com individuos identificados")
> text(x,y,labels=Abreviatura,1:N,col=grupos)
> abline(v=0,lty=2); abline(h=0, lty=2)
> legend(0.02,0,c("Animalia","Monera","Fungi","Plantae", "Protista"),
+ col=c(1,3,2,4,5), text.col="black",bg='gray90',pch=19)
> abline(h=0,lty=2); abline(v=0,lty=2)
```

A.3.4 MDS para os Dados 2

Para desenvolver os cálculos necessários desta secção foram consultadas as seguintes referências bibliográficas: (hel [3]) e (Everitt and Hothorn [21]).

```
> delta<-as.matrix(dist(X),123,123)
> loc<-cmdscale(delta)
> x<--loc[,1]
> y<-loc[,2]
> plot(x, y, type="n", xlab="Coordenadas1",
+ ylab="Coordenadas 2",ylim=c(-2000,5000))
> title("Multidimensional Scaling de cmdscale(delta)")
> text(x,y,labels=Abreviatura,1:N,col=grupos)
> legend(1500,5000,c("Animalia","Monera","Fungi","Plantae", "Protista"),
+ col=c(1,3,2,4,5), text.col="black",bg='gray90',pch=19)
> abline(v=0,lty=2); abline(h=0, lty=2)
> #Diagrama de Shepard
> deltaSh<-Shepard(delta[lower.tri(delta)],delta.mds$points)
> plot(deltaSh,pch=".",xlab="Dissimilhanças",
+ ylab="Distâncias", xlim= range(deltaSh$x),
+ ylim=range(deltaSh$x),col="blue")
> title("Diagrama de Shepard")
> lines(deltaSh$x,deltaSh$yf,type="S")
```

Apêndice B

Tabelas

Na tabela seguinte está apresentada a lista das espécies estudadas que foi estudada na secção 3.2.2.

Tabela B.1: Lista das 123 espécies em estudo, com a designação de cada espécie.

Reino	Espécies	Abreviatura
Animalia	*Bos taurus*	Bt
Animalia	*Callithrix jacchus*	Cj
Animalia	*Cavia porcellus*	Cp
Animalia	*Ciona intestinalis*	Ci
Animalia	*Danio rerio*	Dr
Animalia	*Anolis carolinensis*	Ac
Animalia	*Canis familiaris*	Cf
Animalia	*Choloepus hoffmanni*	Ch
Animalia	*Ciona savignyi*	Cs
Animalia	*Dasypus novemcinctus*	Dn
Animalia	*Dipodomys ordii*	Do
Animalia	*Echinops telfairi*	Et
Animalia	*Equus caballus*	Eq
Animalia	*Erinaceus europaeus*	Ee
Animalia	*Felis catus*	Fc
Animalia	*Gallus gallus*	Gg
Animalia	*Gasterosteus aculeatus*	Ga
Animalia	*Gorilla gorilla*	Ggo
Animalia	*Homosapiens*	Hs
Animalia	*Loxodonta africana*	La
Animalia	*Macaca mulatta*	Mm
Animalia	*Macropus eugenii*	Me
Animalia	*Meleagris gallopavo*	Mg

Continuação na página seguinte

Tabela B - continuação da página anterior

Animalia	*Microcebus murinus*	Mmu
Animalia	*Monodelphis domestica*	Md
Animalia	*Mus musculus*	Mus
Animalia	*Myotis lucifugus*	Ml
Animalia	*Ochotona princeps*	Op
Animalia	*Ornithorhynchus anatinus*	Ao
Animalia	*Oryctolagus cuniculus*	Oc
Animalia	*Oryzias latipes*	Ol
Animalia	*Otolemur garnettii*	Og
Animalia	*Pan troglodytes*	Pt
Animalia	*Xenopus tropicalis*	Xt
Animalia	*Vicugna pacos.vicPac1.57.cdna.abinitio*	Vp
Animalia	*Tursiops truncatus*	Tt
Animalia	*Tupaia belangeri*	Tb
Animalia	*Tetraodon nigroviridis*	Tn
Animalia	*Tarsius syrichta*	Ts
Animalia	*Takifugu rubripes*	Tr
Animalia	*Taeniopygia guttata*	Tg
Animalia	*Sus scrofa*	Ss
Animalia	*Spermophilus tridecemlineatus*	St
Animalia	*Sorex araneus*	Sa
Animalia	*Rattus norvegicus*	Rn
Animalia	*Pteropus vampyrus*	Pv
Animalia	*Procavia capensis*	Pc
Animalia	*Pongo pygmaeus*	Pp
Animalia	*Aedes aegypti*	Aa
Animalia	*Anopheles gambiae*	Ag
Animalia	*Caenorhabditis brenneri*	Cb
Animalia	*Caenorhabditis briggsae*	Cbi
Animalia	*Caenorhabditis elegans*	Ce
Animalia	*Caenorhabditis japonica*	Cja
Animalia	*Caenorhabditis remanei*	Cr
Animalia	*Culex quinquefasciatus*	Cq
Animalia	*Drosophila ananassae*	Da
Animalia	*Drosophila erecta*	De
Animalia	*Drosophila grimshawi*	Dg
Animalia	*Drosophila melanogaster*	Dm
Animalia	*Drosophila mojavensis*	Dmo

Continuação na página seguinte

Tabela B - continuação da página anterior

Animalia	*Drosophila persimilis*	Dp
Animalia	*Drosophila pseudoobscura*	Dps
Animalia	*Drosophila sechellia*	Ds
Animalia	*Drosophila simulans*	Dsi
Animalia	*Drosophila virilis*	Dv
Animalia	*Drosophila willistoni*	Dw
Animalia	*Drosophila yakuba*	Dy
Animalia	*Ixodes scapularis*	Is
Animalia	*Pediculus humanus*	Ph
Monera	*Acidovorax*	A
Monera	*Acinetobacter baumannii*	Ab
Monera	*Aquifex aeolicus*	Aae
Monera	*Archaeoglobus fulgidus*	Af
Monera	*Bacillus anthracis*	Ba
Monera	*Bacillus cereus 03BB102*	Bc
Monera	*Bacillus cereus AH187*	Bce
Monera	*Desulfovibrio vulgaris*	Dvu
Monera	*Escheriachia coli k 12*	Ec
Monera	*Escheriachia coli SE11*	Eco
Monera	*Francisella tularensis*	Ft
Monera	*Halomicrobium mukohataei*	Hm
Monera	*Helicobacter pylori*	Hp
Monera	*Pseudomonas auruginosa*	Pa
Monera	*Salmonella enterics*	Se
Monera	*Salmonella Paratyph*	Sp
Monera	*Salmonella typhi*	Sty
Monera	*Streptococcus mutans*	Sm
Monera	*Streptococcus pneumoniae 70585*	Spn
Monera	*Streptococcus pneumoniae TIGR4*	Spt
Monera	*Aeropyrum pernix*	Ap
Monera	*Haloarcula marismortui*	Hma
Monera	*Halobacteriom salinarum*	Hsa
Monera	*Halobacterium sp*	Hsp
Monera	*Methanococcus jannaschii*	Mj
Monera	*Methanococcus maripaludis*	Mma
Monera	*Methanococcus vannielii*	Mv
Monera	*Pyrococcus furiosus*	Pf
Monera	*Pyrococcus horikoshii*	Pho
Monera	*Thermococcus kodakaraensis*	Tk

Continuação na página seguinte

Tabela B - continuação da página anterior

Fungi	*Aspergillus clavatus*	Acl
Fungi	*Aspergillus flavus*	Afl
Fungi	*Aspergillus fumigatus*	Afu
Fungi	*Aspergillus nidulans*	An
Fungi	*Aspergillus niger*	Ani
Fungi	*Aspergillus oryzae*	Aor
Fungi	*Aspergillus terreus*	At
Fungi	*Neosartorya fischeri*	Nf
Fungi	*Neurospora crassa*	Nc
Fungi	*Saccharomyces cerevisiae*	Sc
Fungi	*Schizosaccharomyces pombe*	Spo
Plantae	*Arabidopsis lyrata*	Al
Plantae	*Arabidopsis thaliana*	Ath
Plantae	*Brachypodium distachyon*	Bd
Plantae	*Oryza indica*	Oi
Plantae	*Oryza sativa*	Os
Plantae	*Populus trichocarpa*	Ptr
Plantae	*Sorghum bicolor*	Sb
Plantae	*Vitis vinifera*	Vv
Protista	*Dictyostelium discoideum*	Dd
Protista	*Plasmodium falciparum*	Pfa
Protista	*Plasmodium knowlesi*	Pk
Protista	*Plasmodium vivax*	Pv

Bibliografia

[1] Biplot for principal components, Abril 2010. http://127.0.0.1:18366/library/stats/html/biplot.princomp.html.

[2] Biplot of multivariate data, Abril 2010. http://127.0.0.1:18366/library/stats/html/biplot.html.

[3] Classical (metric) multidimensional scaling, Agosto 2010. http://127.0.0.1:18366/library/stats/html/cmdscale.html.

[4] The default scatterplot function, Abril 2010. http://127.0.0.1:18366/library/graphics/html/plot.default.html.

[5] Add points to a plot, Abril 2010. http://127.0.0.1:20365/library/graphics/html/points.html.

[6] Principal components analysis, Abril 2010. http://127.0.0.1:18366/library/stats/html/prcomp.html.

[7] Screeplots, Abril 2010. http://127.0.0.1:18366/library/stats/html/screeplot.html.

[8] Singular value decomposition of a matrix, Abril 2010. http://127.0.0.1:18366/library/base/html/svd.html.

[9] *Spotted Express Microarrays*, Fevereiro 2010. http://www-microarrays.u-strasbg.fr/images/spotted/spottedExpressMicroarrays.jpg.

[10] Microarrays, Janeiro 2010. http://pt.wikipedia.org/wiki/Microarranjo.

[11] U. Alon. Affymetrix array data, Abril 2010. http://www.weizmann.ac.il/mcb/UriAlon.

[12] M. C. P. Alonso. *Caracterización multivariante de la realidad socio-económica de la mujer salmantina con empleo irregular*. PhD thesis, Universidad de Salamanca, Departamento de Estadística, 2008.

[13] J. P. Benzécri. *L' Analyse des Données, L' Analyse des Correspondances*, volume 2. 1973.

[14] J. Blasius, P. H. C. Eilers, and J. Gower. Better biplots. In *Computacional Statistic and Data Analysis*, volume 53, pages 3145–3158. June 2009.

[15] J. Cadima. Apontamentos de estatística multivariada. Mestrado em Matemática-ISA/UTL, 2009-2010.

[] F. Cailliez. The analytical solution of the additive constant problem. *Psychometrika*, 48 (2):305–308, 1983.

[17] M. V. M. da Cunha Jr. Análise multidimensional de dados categóricos: A aplicação das análises de correspondência simples e múltipla em marketing e sua integração com técnicas de análise de dados quantitativos. *Cadernos de Estudos do PPGA/EA/UFRGS*, (16/97), Dezembro 1997.

[18] A. D. da Silva, F. Gramaxo, M. E. Santos, A. F. Mesquita, L. Baldaia, and J. M. Félix. *Terra, Universo de Vida 1.ª Parte . BIOLOGIA*.

[19] A. Dourado, S. Vicente, A. Blazquez, and J. Martin. Analisis hj-biplot de la evolución de la produtividad agraria de la comunidad de castilla y leon a lo largo del quiqueno 1991-1995. *Invest. Agr. : Prod. Prot. Veg.*, 14(3):515–530, 1999.

[20] C. Etienne. colored pca biplot, Abril 2010. `http://www.mail-archive.com/r-help@r-project.org/msg54308.html`.

[21] B. S. Everitt and T. Hothorn. *A Handbook of Statistical Analyses Using R*. Chapman Hall, 2010.

[22] Fisher. Machine learning repository, iris data set, Julho 2010. `http://archive.ics.uci.edu/ml/datasets/Iris`.

[23] J. R. Fonseca and H. T. D. Silva. Emprego da análise multivariada na caracterização de acessos de feijão (*Phaseolus vulgaris* l.). *Revista Brasileira de Sementes*, 19(2):334–340, 1997.

[24] F. Freitas, A. Oliveira, F. Carvalho, P. Zimmer, L. Mattos, and M. Kopp. Análise multivariada de populações de azevém (*Lolium multiflorum* l.) em diferentes regimes de água. *R. bras. Agrociência*, 9(1):17–23, jan-mar 2003.

[25] K. R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

[26] S. Gardner-Lubbe, N. J. le Roux, V. Shah, and S. Patwardhan. Biplot methodology in exploratory analysis of microarray data. *Statistical Analysis and Data Mining*, 2:135–145, 2009.

[27] D. Ghosh and A. M. Chinnaiyan. Mixture modelling of gene expression data from microarray experiments. *Oxford University Press*, 18(2):275–286, 2002.

[28] J. Graffelman. *Contributions to the Multivariate Analysis of Marine Environmental Monitoring Data: Methodological Aspects and Applications*. PhD thesis, Departament of Statistics and Operations Research ,Universitat Politècnica de Catalunya, Barcelona, September 2000.

[29] M. Greenacre. Correspondence analysis of raw data. *Ecological Society of America*, 91 (4):958–963, 2010.

[30] M. Greenacre and T. Hastie. The geometric interpretation of correspondence analysis. *Journal of the American Statistical Association*, 82(398):437–447, June 1987.

[31] M. J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, London, 1984.

[32] R. C. L. Idalino, D. L. Roges, and K. R. Santoro. Análise dos polimorfismos do gene hsp70.1 em três espécies. Technical report, Universidade Federal Rural de Pernambuco, Abril 2010.

[33] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall International Editions, 1992.

[34] J. Landgrebe, G. Welzl, T. Metz, M. van Gaalen, H. Ropers, and a. H. W. Wurst. Molecular characterisation of antidepressant effects int the mouse brain using gene expression profiling. *Journal of Psychiatric Research*, 36:119–129, 2002.

[35] F. Álvarez González. *Algunas aportaciones al Análisis de Datos, utilizando técnicas de representación Multivariante*. PhD thesis, Universidad de Cádiz, Facultad de Ciencias, Departamento de Matemáticas, 1999.

[36] K. J. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press London, 1979.

[37] J. H. I. Martel, A. S. Ferraudo, J. R. Môro, and D. Perecin. Estatística multivariada na discriminação de raças amazônicas de pupunheiras (*Bactris gasipaes* kunth) em manaus (brasil). *Rev. Bras. Frutic, Jaboticabal -SP*, 25(1):115–118, 2003.

[38] O. Nenadic and M. Greenacre. Correspondence analysis in r, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20(3):1–13, 2 2007. URL `http://www.jstatsoft.org/v20/i03`.

[39] O. Nenadic and M. Greenacre. Plotting 2d maps in correspondence analysis, Agosto 2010. `http://127.0.0.1:18366/library/ca/html/plot.ca.html`.

[40] O. Nenadic and M. Greenacre. Simple correspondence analysis, Agosto 2010. `http://127.0.0.1:18366/library/ca/html/ca.html`.

[41] E. Oliveira, C. Pedrosa, and R. Pires. *Da Célula ao Universo - Ciências da Terra e da Vida - 11º ANO*. Texto Editora, Lda Lisboa, 1998.

[42] A. S. Pamplona. Análise de correspondência para dados com estrutura de grupo. Master's thesis, Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica, Departamento de Matemática, Fevereiro 1998.

[43] M. Park, J. W. Lee, J. B. Lee, and S. H. Song. Several biplot methods applied to gene expression data. *Journal of Statistical Planning and Inference*, 138:500–515, 2008.

[44] B. A. Pierce. *Genetica - Um Enfoque Conceitual*. 2004.

[45] C. A. Pineda-Vargas, V. M. Prozesky, W. J. Przybylowicz, and J. E. Mayer. Correspondence analysis evaluation of linear nutrient distribuition in root tips of the tropical forage *Brachiaria brizantha*. *Nuclear Instruments and Methods in Physics Research B*, (181):493–498, 2001.

[46] M. Pinheiro, V. Afreixo, G. Moura, A. Freitas, M. Santos, and J. Oliveira. Statistical, computacional and visualization methodologies to unveil gene primary structure. *Methods Inf Med*, 2:163–168, 2006.

[47] J. M. Santos, R. M. Silva, P. Domingues, F. Amado, and M. A. S. Santos. Genómica funcional em aveiro. *Boletim de Biotecnologia*, pages 13–18, 2001.

[48] V. M. Vairinhos and M. P. Galindo. Biplots pmd - data mining centrada em biplots. apresentação de um protótipo. In *XI Jornadas de Classificação e Análise de Dados*. Associação Portuguesa de Classificação e Análise de Dados, Lisboa, 1 a 3 de Abril 2004.

[49] V. M. Vairinhos and V. Lobo. Um projecto de monitorização do estado de condição e predicção de avarias de bordo. `http://www.isegi.unl.pt/docentes/vlobo/Publicacoes/3_10_mecpab.pdf`.

[50] C. S. F. Vieira. Estudo de variáveis discreta: um contributo ao ensino e à genética. Master's thesis, Universidade de Aveiro, Departamento de Matemática, 2007.

[51] M. T. Villalobos Aguayo. Análise de correspondências e modelos log-lineares um enfoque integrado para análise exploratória de dados categóricos. Master's thesis, Campinas. IMECC/ UNICAMP, 1993.

[52] J. L. V. Villardón. Analisis de componentes principales. Departamento de Estadística. Universidad de Salamanca.

[53] M. P. G. Villardón. Una alternativa de representación simultánea: Hj-biplot. *Qüestió*, 10(1):13–23, 1986.